# AIR$^2$: REPAIRING ONLY WHAT MISLEADS THE POLICY FROM SPARSE OUTCOMES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Dense rewards guide learning well but are often misspecified, leading to reward hacking. Sparse trajectory outcomes are reliable but provide weak guidance. Existing approaches learn rewards from feedback, constrain policy optimization, or refine a proxy through residual correction. We study how sparse outcomes can identify corrections that remove misleading incentives while preserving useful dense guidance. The difficulty is that an outcome only measures the total reward of a trajectory, so a correction can fit every outcome and still favor the wrong policy. We quantify this ambiguity with an ambiguity modulus and use it to bound the regret of the repaired policy. The analysis shows that the ambiguity arises in two ways: errors can cancel out within a trajectory, or lie on behavior the data never covers. Fitting the outcomes more accurately fixes neither. We therefore propose **A**daptive **I**dentification for **R**eward **R**epair (AIR$^2$), which queries outcomes on trajectory prefixes to catch errors that cancel, and collects outcomes from an exploration policy to cover what the data missed. AIR$^2$ also estimates a potential function alongside the correction, so that harmless shaping in the proxy is not mistaken for error. The final correction is frozen and passed to a standard RL learner. On 22 continuous control tasks with constructed reward misspecification, AIR$^2$ instantiated with BAC achieves the highest aggregate normalized recovery score among the evaluated methods. We further apply AIR$^2$ with PPO to three published benchmarks and observe substantial improvements over the baseline methods in both experimental settings.

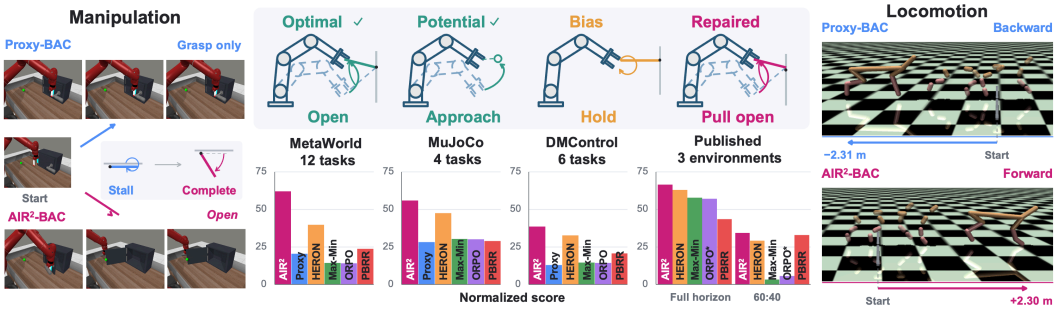



Figure 1: *(Left and right):* **Reward repair in manipulation and locomotion.** The illustrated Proxy-BAC policies stall at grasping the door handle or move backward in HalfCheetah. AIR$^2$-BAC completes door opening and moves forward. *(Top center):* **Reward repair from sparse outcomes.** The decomposition distinguishes the intended reward and potential shaping from bias. AIR$^2$ removes the estimated bias while retaining useful dense guidance. *(Bottom center):* **Comparison of normalized scores.** Bars summarize BAC recovery on the constructed suites and PPO prevention and recovery on the published benchmarks. AIR$^2$ has the highest aggregate score in each displayed comparison.

## 1 INTRODUCTION

Dense proxy rewards are often used when task-level feedback is too sparse to support efficient reinforcement learning. However, a proxy can reward behavior that makes progress toward a task



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

without actually completing it (Pan et al., 2022; Skalse et al., 2022; Gao et al., 2023). In Door-open, rewarding approach and grasp encourages interaction with the handle, but may also favor remaining at the handle rather than opening the door (Figure 1). Such a proxy still contains useful task structure, making it undesirable to discard altogether. We study reward repair from sparse, trusted trajectory outcomes: correcting misleading incentives while retaining useful guidance from the proxy.

A natural approach is to learn an additive correction to the supplied reward (Hatgis-Kessell et al., 2026; Cao et al., 2026). Yet trajectory outcomes supervise accumulated reward rather than individual transitions. Different transition-level corrections could explain the same observations while inducing different policy preferences. Conversely, some reward differences need not be removed. Under suitable boundary conditions, potential shaping changes transition rewards while preserving policy ordering (Ng et al., 1999). This suggests that reward repair is better posed in terms of policy comparisons identified by available outcomes, rather than pointwise recovery of intended reward.

We study this issue from a theoretical lens and characterize when trajectory observations identify the policy comparisons that are relevant for control. We first establish when observationally indistinguishable corrections induce the same policy comparisons over a class of admissible models. We then introduce an **ambiguity modulus** that measures how much approximately outcome-consistent corrections can still disagree in policy value. This turns exact identifiability into a quantitative notion of stability and leads to a regret bound relating observation fit to repaired-policy performance while separating representation, deployment, and policy-optimization errors. The analysis further extends to nonlinear reward models through a local characterization.

The analysis also clarifies what additional feedback can resolve. **Temporal cancellation** occurs when correction errors at different steps offset in the trajectory return. Observations at shorter horizons can expose such differences. **Missing behavioral coverage** arises when relevant errors lie on transitions absent from the observed data, which temporal refinement of the same trajectories cannot recover. We show that additional observations can reduce ambiguity, while errors outside the observed behavior remain unidentified. Since repairing the reward may also shift the policy toward new behavior, feedback collected only under the source policy may be insufficient after repair.

Motivated by these observations, we propose **A**daptive **I**dentification for **R**eward **R**epair (AIR$^2$). AIR$^2$ queries paired continuations from shared states at multiple horizons to expose temporal cancellation, and jointly estimates a transition-level correction and potential to account for shaping contributions. To improve behavioral coverage, it trains an auxiliary policy under the calibrated repaired reward and collects outcomes from the behavior. Candidate corrections are evaluated on both source and exploration behavior before the selected correction is frozen. The resulting scalar reward can be used by standard policy learners without modifying their update rules.

We evaluate AIR$^2$ with BAC (Ji et al., 2024) on 22 continuous-control tasks with constructed reward misspecification, and with PPO (Schulman et al., 2017) on Tomato Watering, Pandemic, and Traffic. On Door-open, AIR$^2$ recovers task completion and improves average return over the strongest baseline by 65.7%. On HalfCheetah, where the proxy assigns additional reward to backward motion, AIR$^2$ restores forward locomotion with a 47.3% improvement. With PPO, AIR$^2$ achieves the highest average intended return on all three prevention tasks and on Tomato Watering recovery. Ablations further examine the roles of potential estimation, temporal feedback, and behavioral coverage.

## 2 Related work

**Reward misspecification and repair.** Inverse reward design infers the designer's intent from a supplied reward while accounting for uncertainty (Hadfield-Menell et al., 2017). ORPO mitigates reward hacking by regularizing occupancy shifts from a reference policy (Laidlaw et al., 2025). MaxMin optimizes against rewards satisfying correlation and moment constraints (Liu et al., 2026). Both rely on proxy correlation under reference behavior. We supervise corrections with sparse trajectory outcomes. HERON learns rewards from a hierarchy of feedback signals (Bukharin et al., 2025). Additive repair and residual reward models reuse prior reward information (Hatgis-Kessell et al., 2026; Cao et al., 2026). We study which corrections cardinal trajectory outcomes identify after accounting for potential shaping, and how the remaining ambiguity affects policy regret.

**Reward identification and feedback.** Reward invariances and comparison metrics formalize the distinction between recovering a reward and preserving its decision consequences (Skalse et al.,

2023; 2024). AIRL (Fu et al., 2018) and Rob-ReL (Lazzati & Metelli, 2026) study reward disentanglement and robust policy comparison, respectively, while RUDDER (Arjona-Medina et al., 2019) redistributes delayed returns for temporal credit assignment. Randomized return decomposition addresses temporal credit assignment by regressing aggregate outcomes over sampled transitions (Ren et al., 2022). We study whether cardinal feedback identifies corrections to a supplied proxy that preserve the intended policy comparisons. Trajectory feedback and active preference learning also study how informative observations support policy learning (Efroni et al., 2021; Sadigh et al., 2017; Lee et al., 2021). Our analysis distinguishes temporal cancellation from missing behavioral coverage and relates both to policy regret through the ambiguity modulus. Further comparisons in additional related work appear in Appendix H.

# 3 Reward repair and identifiability

## 3.1 Problem formulation

Consider an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P$, initial distribution $\rho_0$, and discount $\gamma \in (0,1)$. The learner observes a dense proxy $r^p$ and seeks high return under an intended reward $r^\star$. For a policy $\pi$, write $J^\pi(r) = \mathbb{E}_\pi[\sum_{t\geq 0}\gamma^t r(x_t)]$, where $x_t = (s_t, a_t, s_{t+1}, d_t)$ and $d_t$ denotes true termination. The normalized discounted transition occupancy is $d^\pi$. We compare policies in a declared class $\Pi$ with common initial distribution.

The learner observes transitions and their proxy rewards, supplemented by sparse cardinal outcomes $Y_\tau = G(\tau; r^\star) + \varepsilon_\tau$, where $G(\tau; r) = \sum_{t=0}^{T-1}\gamma^t r(x_t)$ and $\mathbb{E}[\varepsilon_\tau \mid \tau] = 0$. Here, $\tau$ denotes an observed trajectory. Feedback at a specified horizon requires an evaluator that supports that temporal resolution. The learner has no transition-level access to $r^\star$.

To distinguish useful shaping from incentives that can alter policy preferences, we express the proxy as

$$r^p = r^\star + \mathcal{D}_\gamma\Phi^\star + b^\star, \qquad (\mathcal{D}_\gamma\Phi)(x) = \gamma(1-d)\Phi(s') - \Phi(s). \tag{1}$$

Here $\Phi^\star$ is a state potential, and $b^\star$ contains the residual reward error. Separating reward and potential terms also has precedent in AIRL (Fu et al., 2018). We assume absolute discounted integrability of the intended reward and correction. Admissible potentials have integrable initial values and discounted boundary terms that vanish in expectation. Rewards and potentials are zero after true termination. These conditions imply that $J^\pi(r^\star + \mathcal{D}_\gamma\Phi)$ differs from $J^\pi(r^\star)$ by the same constant for every policy (Ng et al., 1999). Shaping therefore preserves policy ordering, whereas the residual may change which policy is preferred.

This invariance determines the correction target. If $(b, \Phi)$ is compatible with the proxy, so is $(b - \mathcal{D}_\gamma\psi, \Phi+\psi)$ whenever this pair is admissible. Subtracting either correction from the proxy produces the same policy ordering. The correction target is therefore an equivalence class defined by policy preferences (Skalse et al., 2023).

Let $f = (b, \Phi)$ belong to a class $\mathcal{F}$ of admissible function pairs. Let $\mu$ denote the sampling distribution over trajectories queried for outcome feedback, induced by the behavior policies and the trajectory selection rule. Summing the potential differences along a trajectory makes the shaping terms telescope, yielding the measurement operator

$$(\mathcal{T}_\mu f)(\tau) = G(\tau; b) - \Phi(s_0) + \gamma^T(1 - d_{T-1})\Phi(s_T). \tag{2}$$

For $f^\star = (b^\star, \Phi^\star)$, the observed discrepancy $y_\tau = G(\tau; r^p) - Y_\tau$ has conditional mean $\mathcal{T}_\mu f^\star$ (Appendix A.2). Each outcome therefore constrains the accumulated correction together with the potential boundary. A trajectory stopped before true termination retains its endpoint term. Jointly modeling the potential accounts for this contribution without including it in the deployed correction.

## 3.2 Identification of policy preferences

Agreement with trajectory outcomes is useful for control only if it constrains the relative values of policies. For a correction difference $h$, define its policy seminorm by

$$\|h\|_\Pi = \sup_{\pi,\pi'\in\Pi} \left|\mathbb{E}_{d^\pi} h - \mathbb{E}_{d^{\pi'}} h\right|. \tag{3}$$

It measures variation in policy advantages in normalized return units. Admissible shaping differences have zero seminorm. Let $\|\cdot\|_\mu$ denote the $L_2(\mu)$ observation norm. For finite ambiguity bounds, we work with a class of finite policy diameter, $D_\Pi(\mathcal{F}) = \sup_{f,\tilde{f}\in\mathcal{F}} \left\|b - \tilde{b}\right\|_\Pi < \infty$. Uniformly bounded correction outputs are sufficient for this condition.

**Proposition 1** (Identification of policy comparisons)**.** *Trajectory measurements identify every policy comparison over $\mathcal{F}$ if and only if, for all $f = (b, \Phi)$ and $\tilde{f} = (\tilde{b}, \tilde{\Phi})$ in $\mathcal{F}$, equality $\mathcal{T}_\mu f = \mathcal{T}_\mu \tilde{f}$ implies $\left\|b - \tilde{b}\right\|_\Pi = 0$.*

Proposition 1 states when observationally equivalent models are interchangeable for policy comparisons (Appendix B.1). The condition concerns differences between admissible pairs, which may exclude directions present in a larger ambient function space. If two models agree on all outcomes but favor opposite choices between two policies, no decision between those policies can be optimal under both models.

The stability of this identification is captured by an *ambiguity modulus*, following the use of moduli of continuity in inverse problems (Donoho, 1994),

$$\omega_\mu(\epsilon) = \sup_{f,\tilde{f}\in\mathcal{F},\ \|\mathcal{T}_\mu(f-\tilde{f})\|_\mu \le \epsilon} \left\|b - \tilde{b}\right\|_\Pi . \tag{4}$$

For fixed $\mathcal{F}$ and $\Pi$, $\omega_\mu(\epsilon)$ is the largest disagreement in normalized policy comparisons between models whose predicted trajectory outcomes differ by at most $\epsilon$ in $L_2(\mu)$. Exact identification of these comparisons corresponds to $\omega_\mu(0) = 0$, even when the reward function is not uniquely identified. If $\omega_\mu(0) > 0$, ambiguity persists with noiseless observations. The diameter condition gives $0 \le \omega_\mu(\epsilon) \le D_\Pi(\mathcal{F})$. At nonzero $\epsilon$, the modulus quantifies the sensitivity of policy comparisons to outcome error. Additional observations restrict the admissible model differences and can therefore improve identification.

**Theorem 2** (Observation refinement)**.** *Appending measurements to $\mathcal{T}_\mu$ with the original weights unchanged gives $\omega_{\mathrm{aug}}(\epsilon) \le \omega_\mu(\epsilon)$ under the direct-sum norm. Admissible correction differences supported outside the observed transitions remain unobserved under temporal refinement.*

Outcomes at additional horizons can expose temporal cancellation, while exploration can reveal errors on unvisited transitions. When the loss is averaged over a different horizon set, its observation weights change, so the numerical tolerance must be rescaled for this comparison (Appendix B.2).

### 3.3 From identification to policy regret

Let $\hat{f} = (\hat{b}, \hat{\Phi}) \in \mathcal{F}$ be a fitted pair and $c$ the correction actually deployed. Let $m_\mu(\tau) = \mathbb{E}[y_\tau \mid \tau]$ be the conditional discrepancy mean. Choose an approximant $f^\dagger = (b^\dagger, \Phi^\dagger)$ in $\mathcal{F}$. Denote population fit and observation misspecification by $\epsilon_{\mathrm{fit}} = \|\mathcal{T}_\mu \hat{f} - m_\mu\|_\mu$ and $\epsilon_{\mathrm{mis}} = \|\mathcal{T}_\mu f^\dagger - m_\mu\|_\mu$. The representation and deployment errors are $\epsilon_{\mathrm{rep}} = \left\|b^\star - b^\dagger\right\|_\Pi$ and $\epsilon_{\mathrm{dep}} = \left\|\hat{b} - c\right\|_\Pi$. The ambiguity modulus connects these errors to the performance of a policy optimized under the repaired reward.

**Theorem 3** (Transfer to control)**.** *Suppose $\hat{\pi} \in \Pi$ is $\xi$-optimal over $\Pi$ for $r^p - c$. Under the conditions of Section 3,*

$$\mathrm{Reg}_{r^\star}(\hat{\pi}) \le \xi + \frac{\omega_\mu(\epsilon_{\mathrm{fit}} + \epsilon_{\mathrm{mis}}) + \epsilon_{\mathrm{rep}} + \epsilon_{\mathrm{dep}}}{1-\gamma}, \tag{5}$$

*where* $\mathrm{Reg}_{r^\star}(\hat{\pi}) = \sup_{\pi\in\Pi} J^\pi(r^\star) - J^{\hat{\pi}}(r^\star)$.

The bound follows by comparing the learned policy with any policy in $\Pi$. Potential shaping contributes the same value offset to both, leaving a policy contrast of the remaining correction error. By the triangle inequality, the observation discrepancy between $\hat{f}$ and $f^\dagger$ is at most $\epsilon_{\mathrm{fit}} + \epsilon_{\mathrm{mis}}$. Their effect on any policy comparison is therefore controlled by the ambiguity modulus. Taking the supremum over comparison policies gives the regret bound (full proof in Appendix C.1).

Eq. (5) separates outcome accuracy from the information needed to make that accuracy useful for control. Better fitting reduces $\epsilon_{\mathrm{fit}}$, while informative observations can reduce $\omega_\mu$. Model approximation affects both $\epsilon_{\mathrm{rep}}$ and $\epsilon_{\mathrm{mis}}$, since the correction and its potential boundary must jointly explain

the outcomes. Deployment error accounts for changes made to the fitted correction, and $\xi$ captures error in policy optimization. Evaluating a calibrated correction at its deployment amplitude includes the scaling in the fitted pair and sets $\epsilon_{\text{dep}} = 0$, provided that pair belongs to $\mathcal{F}$.

A stronger stability guarantee is available when policy contrasts can be reconstructed from trajectory measurements with bounded sensitivity. For a bounded linear operator between Hilbert spaces, a policy contrast represented by $g$ satisfies $|\langle g, f\rangle| \le \kappa \|\mathcal{T}_\mu f\|_\mu$ for all $f$ and some finite $\kappa$ precisely when it admits a dual representation $\mathcal{T}_\mu^* w = g$. The smallest admissible constant is $\min\{\|w\|_\mu : \mathcal{T}_\mu^* w = g\}$. The dual weights express the policy contrast through observed measurements while eliminating the nuisance potential. Large weights amplify observation noise, making the contrast sensitive to small outcome errors even when it is identifiable.

For a nonlinear reward class $\mathcal{F}_R$ within a parameter neighborhood of radius $R$, suppose the linearized measurement map controls linearized correction differences with constant $\kappa$ in the observation and policy norms. If the measurement and correction maps have curvature bounds $L_\mathcal{T}$ and $L_b$, then

$$\omega_{\mu,\mathcal{F}_R}(\epsilon) \le \kappa\epsilon + (\kappa L_\mathcal{T} + L_b)R^2. \tag{6}$$

The $R^2$ term bounds linearization error uniformly over the neighborhood and vanishes for affine parameterizations (Appendix C.3).

## 4 Adaptive Identification for Reward Repair

The regret bound in Eq. (5) guides AIR$^2$ by distinguishing fitting error $\epsilon_{\text{fit}}$ from the ambiguity modulus $\omega_\mu$. We instantiate the method with BAC and PPO as AIR$^2$-BAC and AIR$^2$-PPO. The correction is fitted jointly with a potential function to account for shaping in the proxy. Temporal feedback and policy exploration address the remaining ambiguity. Figure 2 summarizes how AIR$^2$ is integrated with reinforcement learning algorithms.

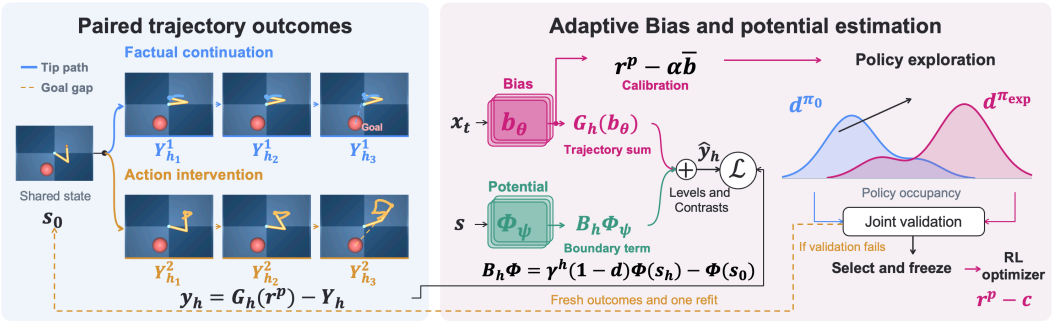



Figure 2: *(Left):* **Paired outcomes at multiple horizons.** Two continuations from a common state provide outcome levels and contrasts. *(Right):* **Adaptive reward repair.** The correction and potential are fitted jointly. Calibration and feedback from an exploration policy guide selection, with one refit if validation fails. The selected correction is frozen for policy training.

### 4.1 Estimating Corrections with a Nuisance Potential

For $f = (b, \Phi)$, let $y_\tau = G(\tau; r^p) - Y_\tau$ and $e_\tau(f) = y_\tau - (\mathcal{T}_\mu f)(\tau)$. Its conditional mean determines the population fitting error, $\epsilon_{\text{fit}} = \|\mathbb{E}[e_\tau(\hat{f}) \mid \tau]\|_\mu$. We fit an ensemble of nonlinear function pairs with mean $\bar{f} = (\bar{b}, \bar{\Phi})$. The objective combines robust losses on outcome levels and matched contrasts with the largest average loss across behavior groups. Outcome levels constrain discrepancies shared by both branches, which cancel in their contrast. The contrast measures relative error between continuations, while the group term prevents a large error under one behavior from being obscured by averaging over the pooled data. Regularization controls correction magnitude and temporal variation, with a separate penalty on the potential (Appendix E.1).

Paired continuations share an initial state, a design also used in trajectory preference learning (Zhu et al., 2023). For observed lengths $T_1, T_2$, write $B_j\Phi = \gamma^{T_j}(1 - d^{(j)}_{T_j-1})\Phi(s^{(j)}_{T_j})$. The initial potential

cancels in their measurement contrast,

$$\Delta\mathcal{T}f = G(\tau^{(2)}; b) - G(\tau^{(1)}; b) + (B_2 - B_1)\Phi. \tag{7}$$

Nonterminal endpoints retain the boundary terms $B_j\Phi$, which the joint model must explain. Omitting them would attribute potential shaping to the correction $b$, although shaping preserves policy comparisons under the boundary conditions of Section 3. Only $b$ is subtracted from the proxy $r^p$, while $\Phi$ accounts for boundary variation in the observations. The correction $b_\theta(x)$ uses transition features, whereas the potential $\Phi_\eta(s)$ uses states. Equation (7) remains valid when both functions are parameterized by nonlinear neural networks.

### 4.2 Temporal observations and calibrated correction

Prefix outcomes expose temporal cancellation through partial returns with their own potential boundaries. By Theorem 2, appending measurements can reduce $\omega_\mu$ when the original weights are retained. Comparisons across horizon sets therefore require rescaling the tolerance of the averaged risk below. For pair $i$, let $\mathcal{H}_i$ be its multiset of recorded horizon slots, $\boldsymbol{e}_{i,h}(f)$ the two branch residuals, and $\Delta e_{i,h}(f) = e^{(2)}_{i,h}(f) - e^{(1)}_{i,h}(f)$ their difference. We assess candidates using the observation risk

$$Q_i(f) = \frac{1}{|\mathcal{H}_i|} \sum_{h \in \mathcal{H}_i} \left[ \frac{\lambda_{\mathrm{obs}}}{2} \left\| \frac{\boldsymbol{e}_{i,h}(f)}{s_{\mathrm{lev},h}} \right\|_2^2 + \left( \frac{\Delta e_{i,h}(f)}{s_{\mathrm{gap},h}} \right)^2 \right]. \tag{8}$$

The scales are fixed from the initial fitting data and shared across candidates, collection rounds and behavior distributions, so their risks are comparable. Resampling treats each trajectory pair and all its prefixes as one block, preserving their statistical dependence. The multiset $\mathcal{H}_i$ retains repeated terminal endpoints when a trajectory ends before a requested horizon and these slots do not constitute independent outcomes.

We use outcomes reserved for calibration to assess the correction at its deployment amplitude. This addresses uncertainty in how strongly the fitted discrepancy should modify the proxy. For $f_\alpha = (\alpha\bar{b}, \bar{\Phi})$ and a finite $\Lambda \subset (0, \infty)$, it selects

$$\widehat{\alpha} \in \arg\min_{\alpha \in \Lambda} \left\{ \mathrm{Quantile}^{\mathrm{boot}}_{1-\delta_{\mathrm{cal}}} \left[ \widehat{\mathbb{E}}_{i \in \mathcal{D}_{\mathrm{cal}}} Q_i(f_\alpha) \right] + \lambda_{\mathrm{cal}} \widehat{\mathcal{P}}_{\mathrm{cal}}(f_\alpha) \right\}. \tag{9}$$

Here $\widehat{\mathcal{P}}_{\mathrm{cal}}$ penalizes the squared change in observation coordinates relative to $f^0 = (0, \bar{\Phi})$. Holding the potential fixed makes this a penalty on the correction alone. The evaluated correction $\alpha\bar{b}$ is exactly the deployed function, so $\epsilon_{\mathrm{dep}} = 0$ when the scaled pair is included in $\mathcal{F}$. The penalty and block bootstrap are specified in Appendix E.2.

### 4.3 Policy exploration and robust selection

Temporal refinement cannot resolve errors on unvisited transitions, so $\omega_{\mu_0}(0)$ may remain positive. An auxiliary policy trained under the calibrated reward supplies feedback beyond the source behavior. Let $\mathcal{V} = \{\mu_0, \mu_{\mathrm{exp}}\}$ contain the sampling distributions induced by the source and exploration policies, with $R_\nu(f) = \mathbb{E}_\nu Q_i(f)$. The criterion $\max_{\nu \in \mathcal{V}} R_\nu(f)$ controls the largest risk over mixtures of these observed behaviors. Source outcomes retain a reference for the original behavior, while exploration outcomes test predictions where the repaired reward directs the learner.

Algorithm 1 evaluates the calibrated correction on fresh outcomes from $\mu_0$ and $\mu_{\mathrm{exp}}$ to assess its predictive accuracy under the behavior induced by reward repair. If validation fails, the estimator is refitted once on augmented fitting data, with separate observations reserved for recalibration and comparison. The model class and exploration policy remain fixed so that both candidates are evaluated under the same behavior distributions.

Let $U_\nu(f)$ be the upper bootstrap quantile of block risk and $L_\nu(f)$ the lower quantile of paired improvement over $f^0 = (0, \Phi_f)$. Eligibility requires $L_\nu(f) > 0$ for every behavior and $\widehat{R}_{\nu,h}(f) \leq (1 + \tau_{\mathrm{val}})\widehat{R}_{\nu,h}(f^0)$ at every horizon, where $\tau_{\mathrm{val}} \geq 0$ is the allowed relative increase in empirical risk. This prevents aggregate improvement from masking excessive error at an individual horizon. With candidate family $\mathcal{C}$ and eligible subset $\mathcal{E}$, selection is

$$\widehat{f} \in \arg\min_{f \in \mathcal{C}_{\mathrm{sel}}} \max_{\nu \in \mathcal{V}} U_\nu(f), \qquad \mathcal{C}_{\mathrm{sel}} = \begin{cases} \mathcal{E}, & \mathcal{E} \neq \varnothing, \\ \mathcal{C}, & \mathcal{E} = \varnothing. \end{cases} \tag{10}$$

Only active corrections enter $\mathcal{C}$. The model with no correction is used solely to test outcome fit. If no candidate is eligible, the selected correction retains its failed validation status. For the weighted block observations, let $\omega_{\mathcal{V}}$ denote the ambiguity modulus under the maximum population $L_2$ norm over $\mathcal{V}$. Under uniform risk control, selection bounds $\epsilon_{\text{fit}}$ in this norm. Corollary 11 combines this error bound with Theorem 3 to control regret through $\omega_{\mathcal{V}}$ (Appendix D.4). The complete procedure is given in Appendix E.2.

**Algorithm 1** **A**daptive **I**dentification for **R**eward **R**epair (AIR$^2$)

**Require:** Source learner state $s_0$, proxy $r^p$, query horizons $\mathcal{H}$

▶ IDENTIFY ON SOURCE BEHAVIOR

Collect separate fitting and calibration blocks from source policy $\pi_0$. ▷ *Eq.* (7)

Fit $f_0$ and calibrate its correction amplitude. ▷ *Eq.* (9)

▶ EXPLORE BEYOND THE SOURCE POLICY

Train auxiliary $\pi_{\text{exp}}$ from $s_0$ under the calibrated reward. ▷ *reward* $r^p - \widehat{\alpha}\bar{b}$

Validate on fresh outcome blocks from $\pi_0$ and $\pi_{\text{exp}}$. ▷ *risk* $Q_i$, *Eq.* (8)

**if** the initial candidate fails validation **then** ▷ *at most one refit*

    Collect fresh feedback and refit once on the augmented fitting set.

    Recalibrate and validate both candidates; keep $\pi_{\text{exp}}$ fixed.

**end if**

▶ SELECT AND DEPLOY

Select $\hat{f}$ and retain its validation status. ▷ *Eq.* (10)

Freeze $c \leftarrow \hat{b}$ and resume $s_0$ with reward $r^p - c$. ▷ *potential* $\hat{\Phi}$ *is not subtracted*

### 4.4 COMPATIBILITY OF AIR$^2$ WITH DIFFERENT REINFORCEMENT LEARNING ALGORITHMS

AIR$^2$ separates reward identification from policy optimization by supplying the learner with the corrected transition reward $r^p(x) - c(x)$. Once $c$ is fixed, the learner retains its original update rules and policy parameterization. This formulation accommodates algorithms that reuse stored transitions as well as those that learn from current policy rollouts. For BAC, the correction is applied to rewards sampled from the replay buffer before value updates (Ji et al., 2024). For PPO, corrected rollout rewards determine returns and advantage estimates used in the standard policy objective (Schulman et al., 2017). Theorem 3 applies to any resulting policy with optimization error at most $\xi$ under the repaired reward, making the regret analysis independent of a particular update rule. More generally, AIR$^2$ can be integrated with learners that optimize numerical rewards, provided that the required transition information and outcome feedback are available.

## 5 EXPERIMENTS

Our experiments address three questions. (1) Can AIR$^2$ recover intended behavior across diverse proxy errors? (2) Does reward repair remain effective with different policy learners in both prevention and recovery? (3) How do potential estimation and feedback design affect recovery?

### 5.1 EVALUATION ON VARIOUS BENCHMARK SUITES

We evaluate AIR$^2$ on 22 constructed tasks from MuJoCo, DeepMind Control and MetaWorld (Todorov et al., 2012; Tassa et al., 2018; Yu et al., 2020). The proxies induce reward hacking by omitting objectives or spatial coordinates, allowing incomplete task execution to receive high rewards. Other constructions distort targets or measurements of motion and control costs, potentially rewarding movement away from the goal or excessive effort. The dynamics and intended objectives remain unchanged. Construction details are provided in Appendix F. The PPO study uses Tomato Watering, a gridworld benchmark, alongside Pandemic and Traffic for epidemic control and traffic management (Pan et al., 2022; Laidlaw et al., 2025; Liu et al., 2026).

**Baselines.** We compare with four methods. (1) PBRR (Hatgis-Kessell et al., 2026) learns additive proxy corrections from preference feedback. (2) HERON (Bukharin et al., 2025) learns rewards from trajectory preferences induced by a feedback hierarchy. (3) Max-Min (Liu et al., 2026) maximizes the minimum return over rewards satisfying correlation and moment constraints. (4) ORPO

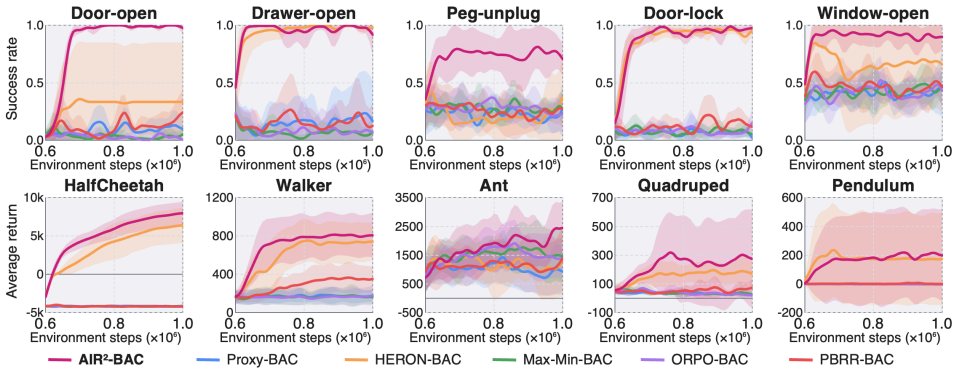



Figure 3: Recovery with BAC. Manipulation success (top) and intended return (bottom).

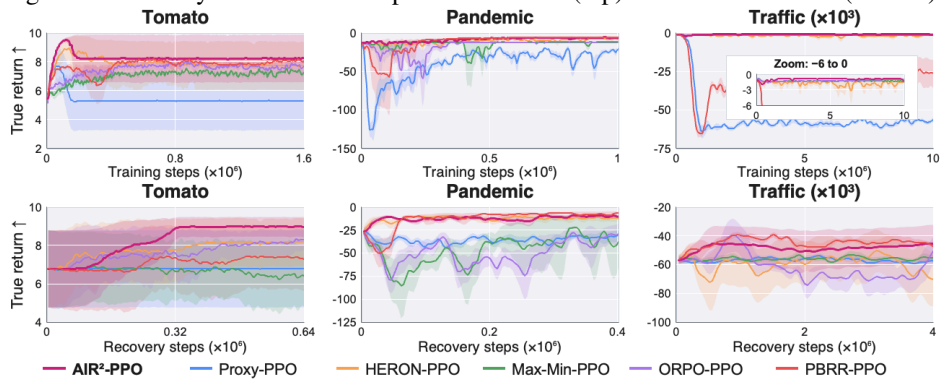



Figure 4: Intended returns with PPO under prevention (top) and recovery (bottom).

(Laidlaw et al., 2025) regularizes occupancy divergence from a reference policy. Proxy optimizes the uncorrected reward. We report average return over training and manipulation success rates.

**Tabletop manipulation.** In Figure 3, AIR$^2$ outperforms the baselines greatly in the speed of convergence and success rate. AIR$^2$ rapidly approaches full success on Door-open, where proxy optimization favors handle interaction without opening the door. Its average recovery return exceeds the strongest repair baseline by 65.7%. AIR$^2$ also maintains higher success on Peg-unplug and Window-open, where the proxies reward initial peg extraction or insufficient window opening. These gains indicate that sparse outcome feedback helps distinguish partial progress from successful task completion. HERON remains competitive on Drawer-open and Door-lock.

**Locomotion and control.** On HalfCheetah, AIR$^2$ recovers forward locomotion despite a proxy that rewards backward motion, improving average recovery return by 47.3% over the strongest competing repair method. Gains on Walker and Quadruped extend to proxies that omit posture objectives, while Ant improves under underestimated control costs. Recovery across these tasks supports the applicability of the correction to both omitted objectives and distorted measurements, beyond the directional bias in HalfCheetah. HERON is comparable on Pendulum, and PBRR leads on Swimmer (Appendix G). Across both task categories, AIR$^2$ achieves the highest aggregate scores on all three benchmark suites in Figure 1.

**Prevention and recovery with PPO.** In Figure 4, AIR$^2$ achieves the highest average return across all three prevention tasks. On Tomato Watering, it also leads during recovery, improving average return over PBRR by 6.8% in prevention and 16.7% in recovery. The advantage is particularly pronounced in Traffic prevention, where AIR$^2$ maintains high intended return throughout training. Its average return is $-0.97 \times 10^3$, compared with $-31.61 \times 10^3$ for PBRR. On Pandemic, AIR$^2$ avoids the pronounced initial degradation observed under proxy optimization. Recovery further demonstrates its ability to improve policies already trained on the misspecified proxy, with higher average return than Proxy on all three tasks (Appendix G.3). Together with the BAC results, these

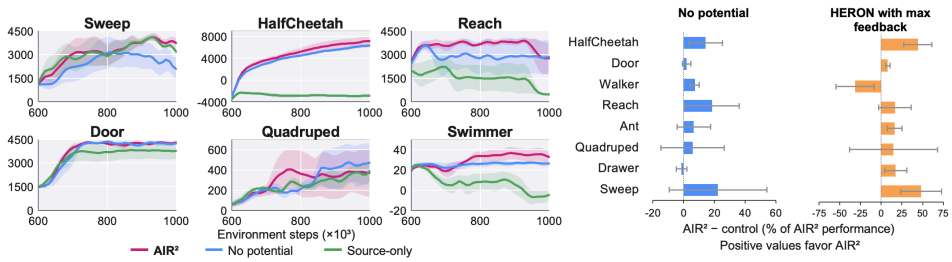



Figure 5: Potential and coverage ablations. Recovery curves (left) and relative return gains over the variant without a potential and HERON with all repair feedback (right).

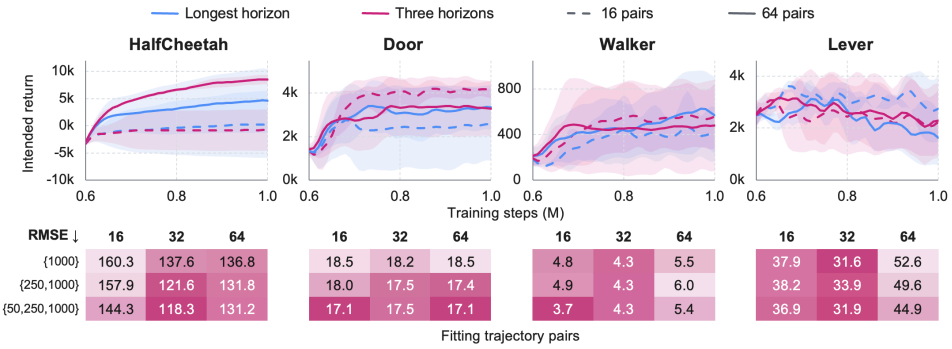



Figure 6: Temporal feedback. Policy recovery (top) and paired outcome prediction error (bottom) with different trajectory counts and horizons.

findings support the applicability of AIR$^2$ across policy learners, both for preventing reward hacking and for recovering intended behavior after proxy optimization.

### 5.2 Ablation Studies

**Potential estimation and behavioral coverage.** Removing the potential while retaining the fitting data produces the largest losses in average return on HalfCheetah and manipulation tasks in Figure 5. Replacing exploration trajectories with an equal number of source policy trajectories also weakens recovery on HalfCheetah and Reach. Moreover, AIR$^2$ retains higher average return than HERON with all stored repair feedback on most tasks. These comparisons support accounting for potential shaping and expanding behavioral coverage beyond simply acquiring more feedback.

**Temporal feedback and policy recovery.** Figure 6 varies temporal observations while restricting fitting data to the source policy. With 64 fitting pairs, three horizons reduce prediction error on all four tasks and increase HalfCheetah's average recovery return from 2947.9 to 5863.3. Policy gains nevertheless vary across tasks, with Walker showing a slight decrease. The distinction between prediction and recovery reflects the separate roles of fit and ambiguity in Theorem 3. Additional horizons also increase outcome queries, so these results measure the joint effect of temporal information and feedback availability (Appendix G.2).

## 6 Conclusion

Dense rewards guide learning but can induce reward hacking when misspecified, whereas sparse trajectory outcomes are reliable but provide limited guidance. AIR$^2$ uses these outcomes to correct misleading incentives while preserving useful dense guidance. Our analysis of identification ambiguity motivates its adaptive feedback collection. On 22 constructed tasks, AIR$^2$ with BAC achieves the highest aggregate normalized recovery scores across all three benchmark suites. Experiments with PPO on three published benchmarks further demonstrate substantial improvements in both prevention and recovery.

### Reproducibility statement

We provide the theoretical and experimental details supporting reproducibility. Section 5 describes the benchmark suites and evaluation settings. Appendices A–D state the assumptions and provide the proofs. Appendix E specifies the model architectures and training procedures, including calibration and integration with policy learners. Appendix F details the construction of tasks with reward misspecification. Appendix G documents the evaluation procedures and ablation settings, including the score normalization used in Figure 1.

### AI use statement

Generative AI tools assisted with locating literature and implementing code, as well as polishing the text of this paper. We did not use generative AI tools for research ideation. The problem formulation, the algorithm design, and the claims made here are the authors' own. All AI generated content has been carefully reviewed by the authors. The authors remain responsible for the manuscript and the submitted artifact.

## References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004. doi: 10.1145/1015330.1015430.

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V. Kulkarni. Bootstrapped reward shaping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15302–15310, 2025. doi: 10.1609/aaai.v39i15.33679.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13544–13555, 2019.

Erdem Bıyık, Nima Anari, and Dorsa Sadigh. Batch active learning of reward functions from human preferences. *ACM Transactions on Human-Robot Interaction*, 13(2):24:1–24:27, 2024a. doi: 10.1145/3649885.

Erdem Bıyık, Nicolas Huynh, Mykel J. Kochenderfer, and Dorsa Sadigh. Active preference-based Gaussian process regression for reward learning and optimization. *The International Journal of Robotics Research*, 43(5):665–684, 2024b. doi: 10.1177/02783649231208729.

Serena Booth, W. Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: Misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5920–5929, 2023. doi: 10.1609/aaai.v37i5.25733.

Alexander Bukharin, Yixiao Li, Pengcheng He, and Tuo Zhao. Deep reinforcement learning from hierarchical preference design. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5875–5893, 2025.

Chenyang Cao, Miguel Rogel-García, Mohamed Nabail, Xueqian Wang, and Nicholas Rhinehart. Residual reward models: Leveraging prior knowledge for efficient preference-based reinforcement learning in robotics. *arXiv preprint arXiv:2507.00611v2*, 2026. Revised 27 August 2026.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

David L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1): 238–270, 1994. doi: 10.1214/aos/1176325367.

R. G. Douglas. On majorization, factorization, and range inclusion of operators on Hilbert space. *Proceedings of the American Mathematical Society*, 17(2):413–415, 1966. doi: 10.1090/S0002-9939-1966-0203464-1.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R. Srikant. Reinforcement learning with segment feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14598–14647, 2025.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7288–7295, 2021. doi: 10.1609/aaai.v35i8.16895.

Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 4705–4713, 2017. doi: 10.24963/ijcai.2017/656.

Sebastian Farquhar, Vikrant Varma, David Lindner, David Elson, Caleb Biddulph, Ian Goodfellow, and Rohin Shah. MONA: Myopic optimization with non-myopic approval can mitigate multi-step reward hacking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16237–16272, 2025.

Lukas Fluri, Leon Lang, Alessandro Abate, Patrick Forré, David Krueger, and Joar Max Viktor Skalse. The perils of optimizing learned reward functions: Low training error does not guarantee low regret. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17306–17377, 2025.

Grant Collier Forbes, Jianxun Wang, Leonardo Villalobos-Arias, Arnav Jhala, and David Roberts. Action-dependent optimality-preserving reward shaping. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17437–17451, 2025.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations*, 2021.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Stephane Hatgis-Kessell, Logan Mondal Bhamidipaty, and Emma Brunskill. Repairing reward functions with feedback to mitigate reward hacking. *arXiv preprint arXiv:2510.13036v2*, 2026. Revised 29 January 2026.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4565–4573, 2016.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tianying Ji, Yu Luo, Fuchun Sun, Xianyuan Zhan, Jianwei Zhang, and Huazhe Xu. Seizing serendipity: Exploiting the value of past success in off-policy actor-critic. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21672–21718, 2024.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. In *International Conference on Learning Representations*, 2024.

W. Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023. doi: 10.1016/j.artint.2022.103829.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019.

Filippo Lazzati and Alberto Maria Metelli. Robustness in the face of partial identifiability in reward learning. In *International Conference on Learning Representations*, 2026.

Chi-Chang Lee, Zhang-Wei Hong, and Pulkit Agrawal. Going beyond heuristics by imposing policy improvement as a constraint. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4383.

Kimin Lee, Laura M. Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6152–6163, 2021.

Haoxin Lin, Hongqiu Wu, Jiaji Zhang, Yihao Sun, Junyin Ye, and Yang Yu. Episodic return decomposition by difference of implicitly assigned sub-trajectory reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13808–13816, 2024. doi: 10.1609/aaai.v38i12.29287.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. Robust optimization for mitigating reward hacking with correlated proxies. In *International Conference on Learning Representations*, 2026.

Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 663–670, 2000.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999.

Evgenii Opryshko, Umangi Jain, and Igor Gilitschenski. Modification-considering value learning for reward hacking mitigation in RL. *arXiv preprint arXiv:2606.28955*, 2026.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022.

Zhizhou Ren, Ruihan Guo, Yuan Zhou, and Jian Peng. Learning long-term reward redistribution via randomized return decomposition. In *International Conference on Learning Representations*, 2022.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems XIII*, 2017. doi: 10.15607/RSS.2017.XIII.053.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347v2*, 2017.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations*, 2024. Revised manuscript arXiv:2309.15257v3, 12 December 2024.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058, 2023.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388, 2015.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. Causal confusion and reward misidentification in preference-based reward learning. In *International Conference on Learning Representations*, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

Hon Tik Tse, Philip S. Thomas, and Scott Niekum. High-confidence policy improvement from human feedback. *Reinforcement Learning Journal*, 6:1455–1479, 2025.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100. PMLR, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775, 2022.

Xiaoying Zhang, Jean-François Ton, Wei Shen, Hongning Wang, and Yang Liu. Mitigating reward overoptimization via lightweight uncertainty estimation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2596.

Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. Interpretable reward redistribution in reinforcement learning: A causal approach. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zihan Zhang, Yuxin Chen, Jason D. Lee, Simon Shaolei Du, and Ruosong Wang. Minimax optimal regret bound for reinforcement learning with trajectory feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74692–74713, 2025.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067, 2023.

## CONTENTS OF THE APPENDIXES


**A Notation and information model** **17**

A.1 Notation and observation model . . . . . 17

A.2 Boundary conditions and potential invariance . . . . . 17

**B Identification and observation refinement** **18**

B.1 Identification and indistinguishability . . . . . 18

B.2 Observation refinement and temporal cancellation . . . . . 19

B.3 Policy coverage and admissible correction classes . . . . . 19

**C Regret bounds and stable reconstruction for nonlinear rewards** **20**

C.1 Regret bounds for repaired policies . . . . . 20

C.2 Stable reconstruction of policy contrasts . . . . . 20

C.3 Nonlinear models and approximation . . . . . 21

**D Statistical specialization and model selection** **22**

D.1 Confidence bounds for policy contrasts . . . . . 22

D.2 Information from additional feedback . . . . . 23

D.3 Geometry of paired temporal observations . . . . . 23

D.4 Selection and control guarantees . . . . . 24

**E Implementation details** **25**

E.1 Reward models and estimation objective . . . . . 25

E.2 Calibration and adaptive repair . . . . . 26

E.3 Integration with policy learners . . . . . 26

**F Construction of scenarios with reward hacking** **27**

F.1 Shared construction principles . . . . . 27

F.2 Omitted objectives and coordinates . . . . . 29

F.3 Miscalibrated targets . . . . . 29

F.4 Incorrect motion and cost measurements . . . . . 30

F.5 Localized false completion . . . . . 30

F.6 Verification of reward construction . . . . . 30

**G Experiment details and additional results** **31**

G.1 Supplementary recovery curves . . . . . 31

G.2 Mechanism ablations . . . . . 31

G.3 PPO experiments on published benchmarks . . . . . 32

G.4 Environment setup . . . . . 33

G.5 Reward repair along policy trajectories . . . . . 35

G.6 Normalized scores . . . . . 35

**H Additional related work and interpretation** **36**

H.1 Reward misspecification and learning from a supplied reward . . . . . 36

H.2 Reward shaping after the classical invariance result . . . . . 36

H.3 Reward comparison and robust decisions . . . . . 36

H.4 Policy optimization under reward misspecification . . . . . 37

H.5 Preference learning and residual rewards . . . . . 37

H.6 Return decomposition and temporal information . . . . . 37

H.7 Feedback collection, uncertainty, and policy improvement . . . . . 37

# A NOTATION AND INFORMATION MODEL

## A.1 NOTATION AND OBSERVATION MODEL

Table 1 summarizes the notation.

Table 1: Mathematical notation.

| Symbol | Meaning |
|---|---|
| $\mathcal{S}, \mathcal{A}, P, \rho_0, \gamma$ | State space, action space, transition kernel, initial law, discount |
| $x = (s, a, s', d)$ | Transition with indicator of true termination $d$ |
| $\pi, \Pi, d^\pi$ | Policy, comparison class, normalized discounted occupancy |
| $J^\pi(r)$ | Expected discounted return under reward $r$ |
| $G(\tau; r)$ | Discounted reward accumulated within an observed trajectory |
| $Y_\tau, \varepsilon_\tau$ | Trusted aggregate outcome and its noise |
| $b, \Phi, f = (b, \Phi)$ | Correction, potential, and joint function pair |
| $\mathcal{F}$ | Declared class of admissible function pairs |
| $\mathcal{D}_\gamma \Phi$ | Discounted potential difference with terminal masking |
| $\mu, \mathcal{T}_\mu, m_\mu$ | Feedback sampling distribution, measurement operator, conditional discrepancy mean |
| $\|\cdot\|_\mu$ | Observation norm, or the declared weighted direct-sum norm |
| $\|\cdot\|_\Pi$ | Seminorm of variation across policy occupancies |
| $\omega_\mu$ | Ambiguity modulus of a specified function and policy class |
| $\epsilon_{\text{fit}}, \epsilon_{\text{mis}}$ | Population fit and observation misspecification errors |
| $\epsilon_{\text{rep}}, \epsilon_{\text{dep}}$ | Representation and deployment errors in the policy seminorm |
| $\xi$ | Policy optimization error under the deployed reward |
| $\kappa$ | Stability constant for reconstructing a policy contrast |
| $e_\tau(f), W_h, Q_i(f)$ | Outcome residual, temporal score matrix, and risk for a block formed by a trajectory pair |
| $\Lambda, \mathcal{C}, \mathcal{E}$ | Positive amplitude grid, calibrated candidates, and candidates meeting the validation criteria |
| $\mathcal{V}, U_\nu, L_\nu$ | Sampling distributions induced by behavior policies, upper risk quantile, and lower improvement quantile |

The learner observes transitions and proxy rewards together with cardinal trajectory outcomes $Y_\tau$ in the intended reward's units. Intended transition rewards and analytic corruption terms are reserved for evaluation. Collection identifiers record shared initial states and prefixes so that statistical analysis can account for their dependence. Binary success fits this model when it is an unbiased observation of a specified terminal reward. Ordinal comparisons require a separate preference likelihood.

Additional temporal measurements require new outcome queries. Each observed trajectory uses its own start state as the discount origin. Transforming observations, including matched differences, transforms both their operator and covariance. For several observation families, we use $\mathcal{T}f = (\mathcal{T}_{\mu_j} f)_j$ with $\|\mathcal{T}f\|^2 = \sum_j w_j \|\mathcal{T}_{\mu_j} f\|^2_{\mu_j}$ and fixed $w_j > 0$. Observation refinement retains the existing weights.

## A.2 BOUNDARY CONDITIONS AND POTENTIAL INVARIANCE

We use an absorbing extension of the MDP after true termination. Rewards and potentials at the absorbing state are zero. For a record containing $T$ transitions, no true termination occurs before the last recorded transition. The final transition may terminate the episode or end at an observation cutoff. The latter event is a truncation and retains the endpoint potential.

For every policy considered, assume absolute discounted integrability of the intended reward and correction. An admissible potential obeys $\mathbb{E}|\Phi(s_0)| < \infty$ and

$$\lim_{T \to \infty} \mathbb{E}_\pi \left[ \gamma^T (1 - d_{T-1}) |\Phi(s_T)| \right] = 0.$$

Bounded potentials suffice when $\gamma < 1$. These conditions justify passing to infinite discounted values. All policies share the same initial distribution.

The normalized transition occupancy is the probability measure $d^\pi = (1-\gamma)\sum_{t\geq 0}\gamma^t \mathcal{L}_\pi(x_t)$ on the absorbing extension. Hence $J^\pi(r) = (1-\gamma)^{-1}\mathbb{E}_{d^\pi} r$. The policy seminorm compares these expectations and vanishes for functions whose occupancy expectations are identical across $\Pi$.

For a finite horizon, include time in the state and set the terminal potential to zero. An earlier observation cutoff retains its endpoint term.

For a trajectory with no earlier termination,

$$\begin{aligned} G(\tau; \mathcal{D}_\gamma \Phi) &= \sum_{t=0}^{T-1} \gamma^t \{\gamma(1-d_t)\Phi(s_{t+1}) - \Phi(s_t)\} \\ &= -\Phi(s_0) + \gamma^T (1-d_{T-1})\Phi(s_T). \end{aligned} \tag{11}$$

Substituting Eq. (1) and the outcome model gives

$$G(\tau; r^p) - Y_\tau = G(\tau; b^\star) - \Phi^\star(s_0) + \gamma^T(1-d_{T-1})\Phi^\star(s_T) - \varepsilon_\tau.$$

Conditional expectation given the trajectory yields the measurement equation. The same calculation applies to trajectories with different start times, with the discount origin set to each trajectory's first transition. Matched trajectories with the same start state cancel their initial potential terms. Their endpoint terms cancel only under an appropriate termination condition or equality of the endpoint potentials.

**Proposition 4** (Potential invariance)**.** *Under the conditions above, $J^\pi(r + \mathcal{D}_\gamma\Phi) = J^\pi(r) - \mathbb{E}_{\rho_0}\Phi$ for every policy with integrable rewards. In particular, adding $\mathcal{D}_\gamma\Phi$ preserves all policy advantages.*

*Proof.* Taking expectations in Eq. (11), adding the reward sum, and passing to the limit gives the stated identity. The discounted endpoint vanishes by admissibility. The remaining expectation over the initial state is common to every policy. □

For admissible $\psi$, the transformation $\tilde{b} = b - \mathcal{D}_\gamma\psi$, $\tilde{\Phi} = \Phi + \psi$ preserves $b + \mathcal{D}_\gamma\Phi$ and every trajectory measurement. The deployed rewards differ by $\mathcal{D}_\gamma\psi$, so they also preserve all policy advantages.

The policy seminorm also removes equivalences induced by the transition kernel or a restricted policy class. Positive reward scaling changes cardinal outcomes and therefore remains distinguishable in this model.

# B Identification and observation refinement

## B.1 Identification and indistinguishability

To prove Proposition 1, define the reward implied by an admissible pair $f = (b, \Phi)$ as $r_f = r^p - b - \mathcal{D}_\gamma\Phi$. Proposition 4 gives

$$J^\pi(r_f) - J^{\pi'}(r_f) = J^\pi(r^p - b) - J^{\pi'}(r^p - b).$$

Two pairs with equal measurements imply the same policy advantages exactly when $\mathbb{E}_{d^\pi}(b - \tilde{b}) = \mathbb{E}_{d^{\pi'}}(b - \tilde{b})$ for every $\pi, \pi' \in \Pi$. Taking the supremum of absolute differences is equivalent to $\|b - \tilde{b}\|_\Pi = 0$. This proves both directions.

Thus policy comparisons must be constant on each observation fiber $\{f \in \mathcal{F} : \mathcal{T}_\mu f = z\}$. For a nonlinear class, the relevant differences are its secants $f - \tilde{f}$.

A decision lower bound additionally requires observationally compatible rewards to favor different policies.

**Proposition 5** (Opposing decisions)**.** *Consider two admissible worlds with the same joint law of all observations. Restrict the final decision to $\{\pi_1, \pi_2\}$. In the first world suppose $J^{\pi_1}(r^\star) - J^{\pi_2}(r^\star) \geq a > 0$, and in the second suppose the reverse advantage is at least $a$. Every randomized learner using these observations has expected regret at least $a/2$ in one world.*

*Proof.* The identical joint distribution of the observed data implies that the learner selects $\pi_1$ with the same probability $q \in [0, 1]$ in both worlds. Writing $R_1$ and $R_2$ for its expected regrets in the two worlds, we obtain

$$\max\{R_1, R_2\} \geq \max\{(1-q)a, qa\} \geq \frac{(1-q)a + qa}{2} = \frac{a}{2}.$$

□

The same argument covers adaptive queries when the complete observation transcript has the same law in both worlds. A query that distinguishes the worlds removes this premise.

### B.2 Observation refinement and temporal cancellation

To prove Theorem 2, write the augmented operator as $\mathcal{T}_{\mathrm{aug}} f = (\mathcal{T}_\mu f, \mathcal{T}_\nu f)$. Its direct-sum norm satisfies

$$\|\mathcal{T}_{\mathrm{aug}}(f - \tilde{f})\|^2 = \|\mathcal{T}_\mu(f - \tilde{f})\|_\mu^2 + \|\mathcal{T}_\nu(f - \tilde{f})\|_\nu^2 \geq \|\mathcal{T}_\mu(f - \tilde{f})\|_\mu^2.$$

Every pair feasible for the augmented modulus at tolerance $\epsilon$ is therefore feasible for the original modulus at that tolerance. The supremum over the smaller set cannot increase.

For the second statement, let an admissible difference $(h, 0)$ have correction support only on a set of transitions absent from the queried trajectories almost surely. Every queried trajectory has $G(\tau; h) = 0$ because the correction difference vanishes on all recorded transitions. Hence its measurement remains zero under any partition of those trajectories. If $\|h\|_\Pi > 0$, ambiguity in policy comparisons persists.

This deterministic inclusion does not require independent observations. Renormalizing the existing weights instead requires rescaling the tolerance before comparing moduli.

For a concrete example of temporal cancellation, consider two transitions with correction features $z_0 = -\gamma z_1 \neq 0$. Their complete trajectory measurement is zero, whereas an observed prefix can separate the two contributions. With an unknown potential, this requires that the additional measurement cannot be absorbed by its endpoint terms. Repeating the same aggregate measurement reduces noise but leaves its exact null space unchanged.

### B.3 Policy coverage and admissible correction classes

Enlarging the policy class $\Pi$ can only increase the seminorm and ambiguity modulus. Identification among source policies therefore need not extend to policies reached after repair.

Suppose all discounted occupancies in $\Pi$ admit density ratios with respect to a common transition law $\nu$, bounded by $C$. Then $|\mathbb{E}_{d^\pi} h| \leq \sqrt{C}\|h\|_{L_2(\nu)}$, and therefore $\|h\|_\Pi \leq 2\sqrt{C}\|h\|_{L_2(\nu)}$. This condition weakens when future policies visit transitions poorly covered by $\nu$. The observation modulus additionally accounts for recovering transition corrections from trajectory outcomes.

The policy diameter $D_\Pi(\mathcal{F})$ introduced in Section 3 bounds every restricted ambiguity modulus. For example, if $\sup_{(b,\Phi)\in\mathcal{F}} \|b\|_\infty \leq B$, then $D_\Pi(\mathcal{F}) \leq 4B$, since discounted occupancies are probability measures. This condition concerns correction contrasts and permits harmless potential variation. The ensemble correction $\bar{b}$ satisfies $\|\bar{b}\|_\infty \leq M_b$ under the parameterization in Appendix E. Thus each calibrated correction $c_\alpha = \alpha\bar{b}$ obeys $\|c_\alpha\|_\infty \leq (\max\Lambda)M_b$. If $M_b$ varies across fits, its largest value is used to obtain a common bound.

Let $\mathcal{C} \subseteq \mathcal{F}$ be any set containing a true admissible representative $f^\star$. Define its correction diameter by $\mathrm{diam}_\Pi(\mathcal{C}) = \sup_{f,\tilde{f}\in\mathcal{C}} \|b - \tilde{b}\|_\Pi$. For any deployed correction $c$ and any $f \in \mathcal{C}$,

$$\|b^\star - c\|_\Pi \leq \mathrm{diam}_\Pi(\mathcal{C}) + \|b - c\|_\Pi.$$

Taking the infimum over $f \in \mathcal{C}$ gives a bound in terms of the correction's distance from the identified set. When $c$ is a representative inside that set, only the set's decision diameter remains.

# C Regret bounds and stable reconstruction for nonlinear rewards

## C.1 Regret bounds for repaired policies

**Proposition 6** (Correction error and regret). *If $\hat{\pi}$ is $\xi$-optimal for $r^p - c$ over $\Pi$, then*

$$\mathrm{Reg}_{r^\star}(\hat{\pi}) \le \xi + \frac{\|b^\star - c\|_\Pi}{1-\gamma}.$$

*Proof.* Let $e = b^\star - c$. By the decomposition, $r^p - c = r^\star + \mathcal{D}_\gamma \Phi^\star + e$. For any $\pi \in \Pi$, potential invariance gives

$$\begin{aligned} J^\pi(r^\star) - J^{\hat{\pi}}(r^\star) &= J^\pi(r^p - c) - J^{\hat{\pi}}(r^p - c) - \{J^\pi(e) - J^{\hat{\pi}}(e)\} \\ &\le \xi + \frac{|\mathbb{E}_{d^\pi} e - \mathbb{E}_{d^{\hat{\pi}}} e|}{1-\gamma} \\ &\le \xi + \frac{\|e\|_\Pi}{1-\gamma}. \end{aligned}$$

Taking the supremum over $\pi$ proves the result. An optimizing policy need not exist because the inequality holds for every comparison policy. □

To prove Theorem 3, the observation triangle inequality gives

$$\|\mathcal{T}_\mu(\hat{f} - f^\dagger)\|_\mu \le \|\mathcal{T}_\mu \hat{f} - m_\mu\|_\mu + \|\mathcal{T}_\mu f^\dagger - m_\mu\|_\mu = \epsilon_{\mathrm{fit}} + \epsilon_{\mathrm{mis}}.$$

Both pairs belong to $\mathcal{F}$, so the definition of the ambiguity modulus gives $\|\hat{b} - b^\dagger\|_\Pi \le \omega_\mu(\epsilon_{\mathrm{fit}} + \epsilon_{\mathrm{mis}})$. A second triangle inequality yields

$$\|b^\star - c\|_\Pi \le \|b^\star - b^\dagger\|_\Pi + \|b^\dagger - \hat{b}\|_\Pi + \|\hat{b} - c\|_\Pi.$$

Apply Proposition 6 to obtain Eq. (5).

## C.2 Stable reconstruction of policy contrasts

Let $\mathcal{H}$ and $\mathcal{Y}$ be real Hilbert spaces and $\mathcal{T} : \mathcal{H} \to \mathcal{Y}$ a bounded linear measurement operator. An element of $\mathcal{H}$ represents a joint correction and potential perturbation. Let a continuous policy contrast have Riesz representer $g \in \mathcal{H}$, so $\ell(h) = \langle g, h \rangle_\mathcal{H}$. Here, $\ker$ denotes the null space of an operator and $^\perp$ denotes the orthogonal complement of a subspace.

**Theorem 7** (Stable dual representation). *The inequality $|\ell(h)| \le \kappa \|\mathcal{T}h\|_\mathcal{Y}$ for every $h \in \mathcal{H}$ holds for some finite $\kappa$ if and only if $g \in \mathrm{range}(\mathcal{T}^*)$. The smallest such constant equals*

$$\kappa_\ell = \min_{\mathcal{T}^* w = g} \|w\|_\mathcal{Y}.$$

*Exact identification has the weaker characterization*

$$\ker \mathcal{T} \subseteq \ker \ell \quad \Longleftrightarrow \quad g \in \overline{\mathrm{range}(\mathcal{T}^*)},$$

*where the bar denotes closure in the norm of $\mathcal{H}$. The two conditions coincide when the adjoint range is closed, including finite-dimensional linear designs.*

*Proof.* If $\mathcal{T}^* w = g$, then $|\ell(h)| = |\langle w, \mathcal{T}h \rangle| \le \|w\| \|\mathcal{T}h\|$ by Cauchy–Schwarz. This proves sufficiency.

Conversely suppose the inequality holds. Define a functional on $\mathrm{range}(\mathcal{T})$ by $L(\mathcal{T}h) = \ell(h)$. If $\mathcal{T}h = \mathcal{T}\tilde{h}$, the assumed inequality implies $\ell(h - \tilde{h}) = 0$, so $L$ is well-defined. Its operator norm is at most $\kappa$. It extends continuously to $\overline{\mathrm{range}(\mathcal{T})}$. The Riesz representation theorem gives a unique $w_0$ in this closed subspace with $L(y) = \langle w_0, y \rangle$ and $\|w_0\| = \|L\|$. Thus $\mathcal{T}^* w_0 = g$.

Any other solution is $w_0 + v$ for $v \in \ker \mathcal{T}^* = (\overline{\mathrm{range}(\mathcal{T})})^\perp$. Pythagoras shows that $w_0$ has minimum norm. Taking the supremum of $|\ell(h)|/\|\mathcal{T}h\|$ over nonzero denominators identifies $\|L\|$

as the smallest admissible constant. If $\mathcal{T} = 0$, the inequality instead forces $g = 0$, and the smallest constant is zero.

Finally, $(\ker \mathcal{T})^{\perp} = \overline{\mathrm{range}(\mathcal{T}^*)}$. The contrast vanishes on the kernel exactly when its representer belongs to this orthogonal complement. □

This classical range condition (Douglas, 1966) characterizes reconstruction of a policy contrast without recovering every reward parameter. For correction and potential designs $Z, U$, the dual constraints are $Z^\top w = q$ and $U^\top w = 0$. The latter removes the nuisance potential. Observation covariance is incorporated by whitening the operator and norm.

Exact identification can nevertheless be unstable in infinite dimensions. Take $\mathcal{H} = \mathcal{Y} = \ell_2$, $(\mathcal{T}h)_j = h_j/j$, and $g_j = 1/j$. The operator is injective, but $\mathcal{T}^* w = g$ would require $w_j = 1$, which is not in $\ell_2$. Indeed, setting $h_j^{(N)} = j/N$ for $j \le N$ and zero otherwise gives $\ell(h^{(N)}) = 1$ while $\|\mathcal{T}h^{(N)}\| = N^{-1/2} \to 0$. This distinction motivates the ambiguity modulus rather than an assumed global linear stability constant (Donoho, 1994).

### C.3 Nonlinear models and approximation

Let $f_\vartheta = (b_\vartheta, \Phi_\vartheta)$ be a differentiable parameterization on $\Theta_R = \{\vartheta : \|\vartheta - \vartheta_0\| \le R\}$. Write the tangent measurement and correction maps as $J_\mathcal{T} = D(\mathcal{T}_\mu f)_{\vartheta_0}$ and $J_b = Db_{\vartheta_0}$. Assume that for every difference between parameters in $\Theta_R$,

$$\|J_b v\|_\Pi \le \kappa \|J_\mathcal{T} v\|_\mu.$$

Suppose the remainders obey

$$\|\mathcal{T}_\mu(f_\vartheta - f_{\vartheta_0}) - J_\mathcal{T}(\vartheta - \vartheta_0)\|_\mu \le \tfrac{1}{2} L_\mathcal{T} \|\vartheta - \vartheta_0\|^2,$$
$$\|b_\vartheta - b_{\vartheta_0} - J_b(\vartheta - \vartheta_0)\|_\Pi \le \tfrac{1}{2} L_b \|\vartheta - \vartheta_0\|^2.$$

Uniformly bounded second derivatives in these norms are sufficient.

**Proposition 8** (Local bound on policy ambiguity)**.** *For the class $\mathcal{F}_R = \{f_\vartheta : \vartheta \in \Theta_R\}$, its restricted modulus $\omega_{\mu,\mathcal{F}_R}$ satisfies Eq. (6).*

*Proof.* Take $\vartheta, \tilde{\vartheta} \in \Theta_R$ and let $v = \vartheta - \tilde{\vartheta}$. Subtracting the two measurement expansions gives $\|J_\mathcal{T} v\|_\mu \le \|\mathcal{T}_\mu(f_\vartheta - f_{\tilde{\vartheta}})\|_\mu + L_\mathcal{T} R^2$. The correction expansions similarly imply $\|b_\vartheta - b_{\tilde{\vartheta}}\|_\Pi \le \|J_b v\|_\Pi + L_b R^2$. Apply tangent stability and take the supremum over pairs with measurement distance at most $\epsilon$. □

The remainder is uniform over $\Theta_R$ and need not vanish with $\epsilon$ at fixed $R$. It is an upper bound, not a lower bound on regret. An affine readout has zero curvature even when its input features are nonlinear.

Potential approximation also enters the observation model. Write $B_\mu \psi(\tau) = -\psi(s_0) + \gamma^T(1 - d_{T-1})\psi(s_T)$. If an admissible approximant has $b^\dagger = b^\star$, then

$$\epsilon_{\mathrm{mis}} = \|B_\mu(\Phi^\dagger - \Phi^\star)\|_\mu, \qquad \epsilon_{\mathrm{rep}} = 0.$$

A restricted potential class can therefore leave observation discrepancies that affect the regret bound through $\omega_\mu$.

Recovering the original $\Phi^\star$ is unnecessary. If the joint class contains $f^\dagger$ with $b^\dagger = b^\star + \mathcal{D}_\gamma(\Phi^\star - \Phi^\dagger)$ and the potential difference is admissible, telescoping gives $\mathcal{T}_\mu f^\dagger = \mathcal{T}_\mu f^\star$ and $\|b^\star - b^\dagger\|_\Pi = 0$. Both approximation errors then vanish. Enlarging the joint class can improve approximation while increasing ambiguity, so model capacity must be matched by informative observations.

# D STATISTICAL SPECIALIZATION AND MODEL SELECTION

## D.1 CONFIDENCE BOUNDS FOR POLICY CONTRASTS

Suppose the correction and potential have fixed features, $b_\theta(x) = \varphi(x)^\top\theta$ and $\Phi_\eta(s) = \psi(s)^\top\eta$. They are fixed before the estimation sample considered below. For observation $i$, define the trajectory feature $z_i = \sum_{t<T_i}\gamma^t\varphi(x_{it})$ and the boundary feature $u_i = -\psi(s_{i0}) + \gamma^{T_i}(1-d_{i,T_i-1})\psi(s_{iT_i})$. The joint parameter is $\beta = (\theta, \eta)$ and the joint design row is $v_i = (z_i, u_i)$.

The observation equation becomes

$$y_i = v_i^\top\beta^\star + \epsilon_i, \qquad y_i = G(\tau_i; r^p) - Y_{\tau_i}.$$

Here $\epsilon_i$ is the negative evaluator noise. Linearity holds in the fixed readouts, without restricting the underlying dynamics.

Let $\mathcal{G}_{i-1}$ include the selected design $v_i$ before the $i$th outcome noise is revealed. Assume $v_i$ is $\mathcal{G}_{i-1}$-measurable and that the noise satisfies $\mathbb{E}[\exp(t\epsilon_i) \mid \mathcal{G}_{i-1}] \le \exp(\sigma^2t^2/2)$ for every real $t$. Independent trajectories with conditionally centered evaluator noise satisfy this condition under a sub-Gaussian tail bound. Correlated prefixes must first be treated as blocks or transformed with a justified conditional covariance model. Deterministic noiseless aggregate labels have $\sigma = 0$.

For $\lambda > 0$, let $V_n = \lambda I + \sum_{i=1}^n v_iv_i^\top$ and $\hat\beta_n = V_n^{-1}\sum_i v_iy_i$. Assume $\|\beta^\star\|_2 \le S$. The regularization radius $S$ and the feature scales are part of the statistical model.

**Theorem 9** (Joint parameter confidence)**.** *Under the stated assumptions, with probability at least* $1-\delta$*,*

$$\|\hat\beta_n - \beta^\star\|_{V_n} \le \sigma\sqrt{2\log\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\delta}} + \sqrt{\lambda}S =: \beta_n(\delta).$$

*Proof.* Write $M_n = \sum_i v_i\epsilon_i$. The self-normalized martingale inequality of Abbasi-Yadkori et al. (2011) gives

$$\|M_n\|_{V_n^{-1}} \le \sigma\sqrt{2\log\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\delta}}$$

on an event of probability at least $1-\delta$. The ridge normal equations yield $V_n(\hat\beta_n - \beta^\star) = M_n - \lambda\beta^\star$. Taking the $V_n^{-1}$ norm and using $V_n \succeq \lambda I$ gives $\|\hat\beta_n - \beta^\star\|_{V_n} \le \|M_n\|_{V_n^{-1}} + \sqrt{\lambda}\|\beta^\star\|_2$. □

The bound accounts jointly for correction and potential uncertainty. It applies to a fixed representation and does not imply coverage for a learned neural ensemble.

For a policy pair define $q_{\pi,\pi'} = \mathbb{E}_{d^\pi}\varphi - \mathbb{E}_{d^{\pi'}}\varphi$ and its joint representer $\tilde q_{\pi,\pi'} = (q_{\pi,\pi'}, 0)$. Theorem 9 implies simultaneous bounds

$$|q_{\pi,\pi'}^\top(\hat\theta - \theta^\star)| \le \beta_n(\delta)\sqrt{\tilde q_{\pi,\pi'}^\top V_n^{-1}\tilde q_{\pi,\pi'}}.$$

The event is simultaneous for every direction because it is a joint norm bound. If a learned policy's occupancy representer is itself estimated, its estimation error must also be included.

Partition $V_n$ by correction and potential coordinates. The correction block of $V_n^{-1}$ is the inverse of

$$\Sigma_{\theta|\eta} = V_{\theta\theta} - V_{\theta\eta}V_{\eta\eta}^{-1}V_{\eta\theta}.$$

The potential block is invertible because $\lambda > 0$. Thus $\tilde q^\top V_n^{-1}\tilde q = q^\top\Sigma_{\theta|\eta}^{-1}q$. This Schur complement measures correction information after accounting for the unknown potential.

Without ridge regularization, let the design matrix be $[Z, U]$. If $U$ has full column rank, projecting outcomes onto the orthogonal complement of its columns gives residual design $(I - U(U^\top U)^{-1}U^\top)Z$. Its Gram matrix is the unregularized counterpart of the Schur complement. For rank-deficient $U$, the same statement uses the orthogonal projector defined by its pseudoinverse. A correction contrast is identified when its representer lies in the row space of this residual design.

### D.2 Information from additional feedback

Adding one fixed design vector $v$ changes the information matrix to $V^+ = V + vv^\top$. The matrix inversion identity gives

$$\tilde{q}^\top (V^+)^{-1} \tilde{q} = \tilde{q}^\top V^{-1} \tilde{q} - \frac{(\tilde{q}^\top V^{-1} v)^2}{1 + v^\top V^{-1} v}.$$

The reduction depends on alignment with the policy contrast under $V^{-1}$, and vanishes for an orthogonal measurement.

For a new observation block with row matrix $W$, the reduction is

$$\tilde{q}^\top V^{-1} W^\top (I + W V^{-1} W^\top)^{-1} W V^{-1} \tilde{q}.$$

The block is whitened using its observation covariance. For singular covariance, whiten the coordinates with positive variance and retain the remaining coordinates as exact linear constraints. The update applies on the resulting feasible parameter subspace.

Suppose the relevant joint design is $p$ dimensional and satisfies $\|v_i\| \le L$. Assume $\lambda_{\min}(\sum_i v_i v_i^\top) \ge n\zeta$ for some $\zeta > 0$, and let $\sup_{\pi,\pi'} \|\tilde{q}_{\pi,\pi'}\| \le Q$. Then

$$\|\hat{b} - b^\star\|_\Pi \le \frac{Q}{\sqrt{\lambda + n\zeta}} \left[ \sigma \sqrt{p \log\left(1 + \frac{nL^2}{p\lambda}\right) + 2\log(1/\delta)} + \sqrt{\lambda} S \right].$$

The determinant bound follows from the arithmetic–geometric mean inequality applied to the eigenvalues of $V_n$. The directional bound follows from $V_n \succeq (\lambda + n\zeta) I$. Combining with Proposition 6 gives a sufficient rate for reward estimation to control downstream regret.

The excitation condition must hold on the design to which this bound is applied. With exact nuisance invariances, use an explicitly reduced design on the identifiable subspace. Repeated measurements cannot compensate for missing relevant directions.

### D.3 Geometry of paired temporal observations

Consider two continuations from the same restored state $s$. Equation (2) gives $\mathcal{T} f(\tau^{(j)}) = G(\tau^{(j)}; b) - \Phi(s) + B_j \Phi$. Subtracting the first branch from the second proves Eq. (7). The initial potential cancels even when the branches have different observed lengths. The final potential cancels only if $B_1 \Phi = B_2 \Phi$, including the case of two truly terminal endpoints. A truncation at a requested horizon alone does not imply this equality.

Let $v = (1, -1)^\top$ and define

$$W_h = \frac{\lambda_{\text{obs}}}{2 s_{\text{lev},h}^2} I_2 + \frac{v v^\top}{s_{\text{gap},h}^2}.$$

The summand for each horizon in Eq. (8) equals $\boldsymbol{e}_{i,h}(f)^\top W_h \boldsymbol{e}_{i,h}(f)$. For the residual coordinates $u_+ = (e^{(1)} + e^{(2)})/\sqrt{2}$ and $u_- = (e^{(2)} - e^{(1)})/\sqrt{2}$, the score diagonalizes as

$$\boldsymbol{e}^\top W_h \boldsymbol{e} = a_h u_+^2 + \left(a_h + 2 s_{\text{gap},h}^{-2}\right) u_-^2, \qquad a_h = \lambda_{\text{obs}} / (2 s_{\text{lev},h}^2).$$

Positive level weight retains the common mode, which contrasts alone would leave unpenalized. These are loss weights, not inverse covariance estimates, and do not assume independence of branches or prefixes. Theorem 2 retains the original observation weights. Averaging over additional horizons instead requires a tolerance adjustment.

If $m$ original horizons are retained among $m' \ge m$ horizons with the same $W_h$ and block sampling law, normalize the respective operators by $m^{-1/2}$ and $(m')^{-1/2}$. They satisfy $\|\mathcal{T}_{\text{old}}^{\text{avg}} h\| \le \sqrt{m'/m} \|\mathcal{T}_{\text{new}}^{\text{avg}} h\|$. Consequently,

$$\omega_{\text{new}}^{\text{avg}}(\epsilon) \le \omega_{\text{old}}^{\text{avg}}\left(\sqrt{m'/m}\, \epsilon\right).$$

For varying block lengths, use a uniform upper bound on the horizon ratios.

For a block $z$ with horizon multiset $\mathcal{H}_z$, stack the vectors $|\mathcal{H}_z|^{-1/2} W_h^{1/2} \boldsymbol{y}_h$ into $y_W(z)$. Stack the corresponding weighted trajectory predictions into $(\mathcal{T}_{\mathcal{V}} f)_\nu(z)$ under each sampling distribution $\nu$. Then $Q_z(f) = \|y_W(z) - (\mathcal{T}_{\mathcal{V}} f)_\nu(z)\|_2^2$. The linear operator acts on functions. The declared admissible class may be nonlinear and must contain every candidate at its deployment amplitude. On tuples of observation functions, set

$$\|v\|_{\mathcal{V}} = \max_{\nu \in \mathcal{V}} \left(\mathbb{E}_\nu \|v_\nu(z)\|_2^2\right)^{1/2}, \qquad \omega_{\mathcal{V}}(t) = \sup_{\substack{f, \tilde{f} \in \mathcal{F} \\ \|\mathcal{T}_{\mathcal{V}}(f - \tilde{f})\|_{\mathcal{V}} \le t}} \left\| b - \tilde{b} \right\|_\Pi .$$

Norms are taken within each block before averaging, preserving dependence among prefixes. The transfer proof applies to this maximum norm through its triangle inequality.

### D.4 Selection and control guarantees

Condition on the data constructing a finite candidate family $\mathcal{C} \subseteq \mathcal{F}$, the fixed normalization scales, and the sampling distributions over trajectory pairs, before collecting fresh validation blocks. Let $m_{W,\nu}(z) = \mathbb{E}[y_W \mid z, \nu]$, where $z$ includes both trajectory branches. Assume finite second moments and feedback noise with conditional mean zero given the full block. Exact outcomes satisfy this condition. For noisy outcomes, centering is required jointly given both branches.

Write $D_\nu(f)^2 = \mathbb{E}_\nu \|(\mathcal{T}_{\mathcal{V}} f)_\nu - m_{W,\nu}\|_2^2$ and $\sigma_\nu^2 = \mathbb{E}_\nu \|y_W - m_{W,\nu}\|_2^2$. Expanding the squared norm and conditioning on the block eliminates the cross term, giving $R_\nu(f) = D_\nu(f)^2 + \sigma_\nu^2$. All covariance between feedback coordinates is included in $\sigma_\nu^2$. Define

$$\epsilon_{\mathcal{C}} = \min_{f \in \mathcal{C}} \max_\nu D_\nu(f), \qquad \Delta_{\text{noise}} = \max_\nu \sigma_\nu^2 - \min_\nu \sigma_\nu^2.$$

The noise term permits different variances across behavior distributions.

Let $S(f) = \max_\nu R_\nu(f)$ and $\widehat{S}(f) = \max_\nu U_\nu(f)$. For any probability vector $p$ over $\mathcal{V}$, $\sum_\nu p_\nu R_\nu(f) \le S(f)$, with equality at a maximizing distribution. This proves the mixture interpretation in Section 4. It concerns mixtures of the collected behaviors, not all policies in $\Pi$.

With the nonempty selection set $\mathcal{C}_{\text{sel}}$ from Eq. (10), define

$$\Delta_{\text{elig}} = \min_{f \in \mathcal{C}_{\text{sel}}} S(f) - \min_{f \in \mathcal{C}} S(f) \ge 0.$$

The set can depend on validation. When all candidates fail, the rule uses $\mathcal{C}_{\text{sel}} = \mathcal{C}$ and $\Delta_{\text{elig}} = 0$. Let selection satisfy $\widehat{S}(\hat{f}) \le \min_{\mathcal{C}_{\text{sel}}} \widehat{S} + \eta$.

**Proposition 10** (Selection from a finite candidate family)**.** *On the event* $\max_{f,\nu} |U_\nu(f) - R_\nu(f)| \le \rho$*,*

$$\|\mathcal{T}_{\mathcal{V}} \hat{f} - m_W\|_{\mathcal{V}}^2 \le \epsilon_{\mathcal{C}}^2 + \Delta_{\text{elig}} + 2\rho + \eta + \Delta_{\text{noise}}.$$

*Proof.* The maximum is Lipschitz in the sup norm, so $|\widehat{S}(f) - S(f)| \le \rho$ for every candidate. Comparing the selected candidate to a population minimizer over $\mathcal{C}_{\text{sel}}$ gives $S(\hat{f}) \le \min_{\mathcal{C}} S + \Delta_{\text{elig}} + 2\rho + \eta$. Furthermore,

$$\max_\nu D_\nu(\hat{f})^2 \le S(\hat{f}) - \min_\nu \sigma_\nu^2, \qquad \min_{\mathcal{C}} S \le \epsilon_{\mathcal{C}}^2 + \max_\nu \sigma_\nu^2.$$

Combining these inequalities proves the claim. □

Combining this bound with Theorem 3 gives a control guarantee under the norm $\|\cdot\|_{\mathcal{V}}$ defined above. Its ambiguity modulus is $\omega_{\mathcal{V}}$, and misspecification is measured in the same norm.

**Corollary 11** (Selection and control)**.** *Condition on candidate construction and the sampling distributions. Suppose* $\max_{f,\nu} |U_\nu(f) - R_\nu(f)| \le \rho$ *and selection has optimization error* $\eta$*. Under the transfer assumptions, deploying the selected correction exactly gives*

$$\text{Reg}_{r^\star}(\hat{\pi}) \le \xi + \frac{\omega_{\mathcal{V}}\left(\sqrt{\epsilon_{\mathcal{C}}^2 + \Delta_{\text{elig}} + 2\rho + \eta + \Delta_{\text{noise}}} + \epsilon_{\text{mis}}\right) + \epsilon_{\text{rep}}}{1 - \gamma}. \tag{12}$$

*Proof of Corollary 11.* Choose $f^\dagger \in \mathcal{F}$ as in Theorem 3, now with $\epsilon_{\text{mis}} = \|\mathcal{T}_{\mathcal{V}} f^\dagger - m_W\|_{\mathcal{V}}$. The triangle inequality and the proposition bound $\|\mathcal{T}_{\mathcal{V}}(\hat{f} - f^\dagger)\|_{\mathcal{V}}$ by $\sqrt{\epsilon_{\mathcal{C}}^2 + \Delta_{\text{elig}} + 2\rho + \eta + \Delta_{\text{noise}}} + \epsilon_{\text{mis}}$. Apply the defining modulus and then the policy comparison argument in Appendix C.1. The potential contributes the same value offset to both policies. Exact deployment sets $\epsilon_{\text{dep}} = 0$. This proves Eq. (12). □

A sufficient uniform risk bound follows from independent validation blocks. For each behavior, suppose validation contains $n_\nu$ conditionally independent blocks formed by trajectory pairs and $Q_i(f) \in [0, L_\nu]$ almost surely for every candidate. Hoeffding's inequality and a union bound imply

$$\max_{f,\nu} |\widehat{R}_\nu(f) - R_\nu(f)| \le \max_\nu L_\nu \sqrt{\frac{\log(2|\mathcal{C}||\mathcal{V}|/\delta)}{2n_\nu}} =: \rho_{\text{mean}}$$

with conditional probability at least $1-\delta$. For the implemented score, set $B_{\text{boot}} = \max_{f,\nu} |U_\nu(f) - \widehat{R}_\nu(f)|$. A triangle inequality gives $\rho = \rho_{\text{mean}} + B_{\text{boot}}$. This uses the observed bootstrap deviation without assuming nominal quantile coverage. The loss bound must follow from model and outcome bounds. Independence is required across blocks within each behavior, allowing dependence within blocks and across behavior groups.

For adaptive fitting, condition before each fresh validation collection and allocate failure probability across rounds. A union bound then covers stopping and selection. The optional refit uses separate fresh blocks for fitting and calibration, preserving validation independence.

# E Implementation details

## E.1 Reward models and estimation objective

AIR$^2$-BAC uses an ensemble of four models. Each model contains a correction network with hidden dimension 128 and a potential network with hidden dimension 32. The correction input is the standardized concatenation $[s, a, s' - s, d]$. The potential input is the standardized state. Network parameters and input normalizers are initialized afresh for each fit using the declared training data.

The correction is the ensemble mean of bounded signed outputs, scaled by a calibrated positive amplitude. Pointwise uncertainty shrinkage is disabled.

From the training records define the discrepancy scale $s_y = \max\{0.1, \text{median}_\tau |G(\tau; r^p) - Y_\tau|\}$ and proxy scale $s_r = \max\{0.1, \text{RMS}(r^p_{\text{sampled}})\}$. The output ranges use $M_b = \max\{0.25, 1.25 \text{ quantile}_{.99} |G(\tau; r^p) - Y_\tau|\}$ and $M_\Phi = \max\{1, s_r\}$ with signed tanh outputs. These scales are derived from the fitting data and held fixed during training.

For fitting families $g \in \mathcal{G}$, matched pairs $(i, j)$ within a family, and ensemble members $f_k = (b_k, \Phi_k)$ with mean $\bar{f}$, the target is

$$\begin{aligned} \underset{f_{1:K}}{\text{minimize}} \quad & \lambda_{\text{lev}} \widehat{\mathbb{E}}_{k,\tau} \rho_{\text{H}}\left(\frac{e_\tau(f_k)}{s_y}\right) + \widehat{\mathbb{E}}_{k,(i,j)} w_{ij} \rho_{\text{H}}\left(\frac{e_j(f_k) - e_i(f_k)}{s_{\text{gap},g}}\right) \\ & + \lambda_{\text{fam}} \max_{g \in \mathcal{G}} \widehat{\mathbb{E}}_{\tau|g} \rho_{\text{H}}\left(\frac{e_\tau(\bar{f})}{s_y}\right) + \mathcal{R}(f_{1:K}). \end{aligned} \tag{13}$$

The expectations use stratified batches and normalized pair weights. Here $K = 4$ and $\lambda_{\text{lev}} = \lambda_{\text{fam}} = 0.25$. $\rho_{\text{H}}$ denotes smooth $L_1$ loss. Each pair scale is derived from its fitting family. Pair weights are proportional to $1 + 2\min\{|(y_j - y_i)/s_{\text{gap}}|, 4\}$ and normalized within the batch. The family term uses the largest mean level loss across fitting families.

The complete objective is

$$\begin{aligned} \mathcal{L} &= 0.25\mathcal{L}_{\text{level}} + \mathcal{L}_{\text{pair}} + 0.25\mathcal{L}_{\text{family}} + \mathcal{R}_b + 0.001\mathcal{R}_\Phi, \\ \mathcal{R}_b &= \frac{0.001}{s_r}\mathbb{E}_{k,t}|b_k(x_t)| + \frac{0.01}{s_r^2}\mathbb{E}_{k,t}[b_k(x_{t+1}) - b_k(x_t)]^2. \end{aligned} \tag{14}$$

Masks exclude padding and invalid adjacent transitions. $\mathcal{R}_\Phi$ penalizes the squared potential at the initial state, divided by $s_y^2$. It regularizes the nuisance representation separately from the transition correction.

Training takes 4000 optimizer steps with correction and potential learning rates $3 \cdot 10^{-4}$ and $10^{-4}$. The base batch contains 32 stratified trajectories with appended matched pairs. All members use every fitting family with unit weights. Retraining the models does not imply that new feedback data were independently collected.

### E.2 Calibration and adaptive repair

For a deployed candidate $c_\alpha = \alpha\bar{b}$, with associated potential $\bar{\Phi}$, define $e_\tau = y_\tau - \mathcal{T}_\mu(c_\alpha, \bar{\Phi})(\tau)$. For matched branches $i, j$ and horizon $h$, the score coordinates are

$$\left(\sqrt{0.125}\frac{e_i}{s_{\mathrm{level},h}}, \sqrt{0.125}\frac{e_j}{s_{\mathrm{level},h}}, \frac{e_j - e_i}{s_{\mathrm{gap},h}}\right).$$

The squared norm combines level weight $\lambda_{\mathrm{obs}} = 0.25$ with contrast weight 1. All normalizers are fixed from the initial fitting data and shared across candidates, data collection rounds, and behavior distributions. Prefix risks are averaged within each trajectory pair before the pairs are resampled as blocks.

Calibration selects $\alpha \in \{0.25, 0.5, 0.75, 1\}$ by the 97.5% bootstrap upper quantile of mean block risk, plus 0.001 times the mean squared norm of the correction in observation coordinates. It uses 5000 bootstrap draws. In the notation of Eq. (8), this penalty is

$$\widehat{\mathcal{P}}_{\mathrm{cal}}(f) = \widehat{\mathbb{E}}_{i \in \mathcal{D}_{\mathrm{cal}}} \frac{1}{|\mathcal{H}_i|} \sum_{h \in \mathcal{H}_i} [\boldsymbol{e}_{i,h}(f) - \boldsymbol{e}_{i,h}(f^0)]^\top W_h [\boldsymbol{e}_{i,h}(f) - \boldsymbol{e}_{i,h}(f^0)].$$

The shared potential in $f$ and $f^0$ cancels in this penalty. The comparator assesses predictive improvement and is excluded from deployment. Block resampling requires at least 16 trajectory pairs.

Validation uses 20,000 bootstrap draws, resampling trajectory pairs as blocks separately for the source and exploration distributions. The nominal one-sided tail allocation is $0.025/(|\mathcal{C}||\mathcal{V}|)$, with one or two calibrated models and two behavior distributions. Eligibility requires a positive lower bootstrap bound for comparator risk minus candidate risk and mean risk at every horizon at most 1.05 times its comparator, corresponding to $\tau_{\mathrm{val}} = 0.05$. A candidate must satisfy the rule on both distributions.

If the first candidate is ineligible, one additional round collects separate fitting, calibration and validation blocks under the existing behavior distributions. The second model is fitted on the augmented data, then both models are recalibrated and validated on fresh blocks. Selection minimizes the largest upper risk among jointly eligible candidates, or among all active candidates if none is eligible.

Algorithm 2 gives the complete procedure. The saved learner state $s_0$ contains the source policy $\pi_0$, and $\mathcal{V} = \{\mu_0, \mu_{\mathrm{exp}}\}$ denotes the source and exploration trajectory distributions. Fitting, calibration and validation use separate blocks. The candidate model class and exploration policy remain fixed during the optional refit.

### E.3 Integration with policy learners

The policy learner receives the frozen transition reward $r^p(x) - c(x)$ through its ordinary reward interface. Estimator fitting requires transitions and aggregate outcomes together with the initial and final observations needed for the potential boundary.

For AIR$^2$-BAC, continuation restores the source policy and critic together with their optimizer state and replay buffer (Ji et al., 2024). The correction is applied to every sampled replay transition, including those collected before it was frozen. The auxiliary exploration policy supplies feedback without replacing the source learner.

For PPO, corrected rollout rewards are used consistently in returns and advantage estimates (Schulman et al., 2017). With value function $V_\psi$, the temporal difference residual is

$$\delta_t^c = r^p(x_t) - c(x_t) + \gamma(1 - d_t)V_\psi(s_{t+1}) - V_\psi(s_t).$$

The clipped policy objective and ordinary value updates are unchanged. Neither learner differentiates through reward fitting.

**Algorithm 2** **A**daptive **I**dentification **f**or **R**eward **R**epair (AIR$^2$)

**Require:** Source learner state $s_0$, proxy $r^p$, horizons $\mathcal{H}$, amplitudes $\Lambda$

*// Fit the discrepancy and calibrate the correction*

$(\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{cal}}) \leftarrow \text{Collect}(\pi_0, \mathcal{H})$

$f_0 \leftarrow \text{Fit}(\mathcal{D}_{\text{fit}})$ by Eq. (13)

$\tilde{f}_0 \leftarrow \text{Calibrate}(f_0, \mathcal{D}_{\text{cal}})$ by Eq. (9)

*// Explore beyond the source policy distribution*

$\pi_{\text{exp}} \leftarrow \text{BAC}(s_0, r^p - \tilde{b}_0)$

$\mathcal{D}_{\text{val}} \leftarrow \text{Collect}(\mathcal{V}, \mathcal{H})$

$\mathcal{C} \leftarrow \{\tilde{f}_0\}, \quad \mathcal{E} \leftarrow \text{Validate}(\mathcal{C}, \mathcal{D}_{\text{val}})$

*// If validation fails, collect fresh feedback and refit once*

**if** $\mathcal{E} = \varnothing$ **then**

  $(\mathcal{D}^+_{\text{fit}}, \mathcal{D}^+_{\text{cal}}, \mathcal{D}^+_{\text{val}}) \leftarrow \text{Collect}(\mathcal{V}, \mathcal{H})$

  $f_1 \leftarrow \text{Fit}(\mathcal{D}_{\text{fit}} \cup \mathcal{D}^+_{\text{fit}})$

  $\mathcal{C} \leftarrow \{\text{Calibrate}(f_j, \mathcal{D}^+_{\text{cal}})\}_{j=0}^{1}$

  $\mathcal{E} \leftarrow \text{Validate}(\mathcal{C}, \mathcal{D}^+_{\text{val}})$

**end if**

*// Freeze the correction and resume the common source learner*

$\hat{f} \leftarrow \text{Select}(\mathcal{C}, \mathcal{E})$ by Eq. (10)

$c \leftarrow \hat{b}$ ▷ Freeze $c$, without subtracting $\hat{\Phi}$

**return** $\text{BAC}(s_0, r^p - c)$

Adaptive collection requires paired continuations from restored states and outcome queries on their prefixes. Applying the procedure in another domain requires corresponding collection and evaluation access. Theorem 3 applies when the resulting policy satisfies its $\xi$-optimality condition under $r^p - c$. For a regularized or constrained learner, $\xi$ also accounts for the difference between its optimization objective and this reward objective.

# F Construction of scenarios with reward hacking

## F.1 Shared construction principles

Each scenario pairs an unchanged physical task with a specified proxy reward. The native environment supplies the intended reward for evaluation. The proxy modifies a reward measurement, such as direction, target location, or a completion component. Physics, observations, and native success criteria remain those of the registered environment, subject to the declared episode wrapper. MetaWorld uses its native v2 reward implementation and v3 task IDs. MuJoCo tasks use the v5 Gymnasium registrations.

The constructed proxies include various forms of reward misspecification and the shaping function is not provided to the learner. For clarity, write $\tilde{r}$ for the proxy component before the common potential is added. Table 2 specifies our 22 constructed scenarios. The same task ordering applies to every method. The table records the actual mechanism in the source implementation.

Table 2: All 22 constructed scenarios. The middle column gives the reward measurement error and its frozen parameters. The last column describes the suboptimal incentive it introduces. Native success is unchanged.

| | Native task | Proxy construction | Resulting suboptimal incentive |
|---|---|---|---|
| 1 | MetaWorld Door-open | Retain only 2 $\text{Hamacher}(\text{ready}, \text{grasp})$. Omit the native opening/completion reward. | Approach and manipulate the handle without completing door opening. |

Table 2 (continued)

| | Native task | Proxy construction | Resulting suboptimal incentive |
|---|---|---|---|
| 2 | MetaWorld Drawer-open | Evaluate opening at 60% of the native target displacement. Retain the native 2 cm tolerance and approach term. | Stop at a partially opened drawer that the proxy treats as complete. |
| 3 | DMC Ball-in-Cup/catch | Add $2\ell(1-a)$, where $\ell$ and $a$ are height and lateral tolerances scaled by the native capture extents. | Lift the ball to cup height while remaining laterally outside the cup. |
| 4 | MuJoCo HalfCheetah-v5 | Add $10\tanh(\max\{-v_x, 0\}/4)$ to the native reward. | Collect false odometry credit through backward motion. |
| 5 | DMC Walker/walk | Use the native movement factor $(5m+1)/6$ without its posture factor. | Make forward progress with poor standing posture. |
| 6 | MuJoCo Ant-v5 | Replace actuator cost $c$ by $0.1\tanh(c/0.1)$. | Use excessive control effort whose true cost is underreported. |
| 7 | DMC Quadruped/walk | Retain the native forward movement tolerance and omit its upright factor. | Move forward while rolled or insufficiently upright. |
| 8 | MetaWorld Reach-v3 | Compute target distance in $(x, z)$ only. Preserve native radius and three-dimensional initial margin. | Align in the measured plane while remaining incorrect in depth. |
| 9 | MetaWorld Lever-pull | Evaluate completion at 60% of the native $\pi/2$ rotation. Preserve the 4 cm tolerance and grasp term. | Stop at a partially rotated lever. |
| 10 | MetaWorld Faucet-open | Add $12g(1-i)$, with native proximity $g$ and completion score $i$. | Remain near the handle with incomplete opening to collect credit for the remaining error. |
| 11 | DMC Cartpole/swingup | Add $2uav(1-c)$ to native $ucav$, with upright, centering, control, and velocity factors. | Devalue cart centering while preserving the other native factors. |
| 12 | MuJoCo Swimmer-v5 | Add $2.5\tanh(\max\{-v_x, 0\})$. | Exploit directionless odometry through reverse swimming. |
| 13 | DMC Reacher/easy | Add $1.25\sin^2(\pi(d/R-1))$ for $R < d < 2R$, zero elsewhere. | Hold the fingertip just outside the native target radius. |
| 14 | MetaWorld Push-v3 | Omit the lateral $x$ coordinate from the distance between the object and target and proxy completion. Keep grasp and contact branches. | Push to the correct depth while missing the target laterally. |
| 15 | MetaWorld Window-open | Move the measured opening target from 20 cm to 8 cm. Retain the native 5 cm radius and 15 cm margin. | Complete only part of the required window opening. |
| 16 | MetaWorld Sweep-v3 | Add $8(\lvert u\rvert - u)$ for $u = \tanh(\Delta x_{\text{object}}/0.01)$. | Earn extra reward by moving the object away from the rightward goal. |
| 17 | MetaWorld Peg-unplug-side | Keep approach and initial extraction terms. Remove $5i$ and the success override to 10. | Extract the peg slightly and remain near the fixture without completing the unplugging motion. |
| 18 | MetaWorld Door-lock | Shift the measured lock target upward by 5 cm. Preserve the 0.5 cm reward radius and native approach term. | Optimize a displaced locking target while native success remains 2 cm. |

Table 2 (continued)

| | Native task | Proxy construction | Resulting suboptimal incentive |
|---|---|---|---|
| 19 | MetaWorld Coffee-pull | Add $8(\|u\| - u)$ for $u = \tanh((d_t - d_{t+1})/0.01)$. | Move the cup away from its goal to receive false progress credit. |
| 20 | MetaWorld Handle-pull | Add $16g(1 - i)$ using native grasp and completion components. | Maintain grasp while leaving the handle short of completion. |
| 21 | DMC Pendulum/swingup | Add $1.5(\|u\| - u)$ for $u = \tanh((\cos\theta_{t+1} - \cos\theta_t)/0.05)$. | Collect credit while moving away from the upright configuration. |
| 22 | MuJoCo Reacher-v5 | Rotate the measured target by $+90^\circ$ in the target plane. Retain both distance coordinates and the native action cost. | Reach a displaced target while the physical intended target stays fixed. |

### F.2 Omitted objectives and coordinates

Door uses the native readiness and grasp components combined by the Hamacher product. The completion reward is omitted from the proxy component. The physical door remains movable and the intended reward continues to measure opening. Thus the proxy preserves a useful prerequisite skill but provides little incentive to finish the task.

Walker and Quadruped retain movement rewards while dropping posture factors. Their native movement tolerances and target speeds are unchanged. In the wrapper, the added discrepancy is the nonnegative difference between the component that rewards movement alone and the native reward. This construction creates an incentive for locomotion without the desired body configuration.

MetaWorld Reach computes the end-effector offset in $x$ and $z$ and omits $y$. Its tolerance has native radius 0.05 and the native three-dimensional initial margin. Push instead omits $x$ from the distance between the object and target. Push retains the native contact condition, gripper condition, grasp reward, and reward branch structure. Only its measured distance to the target and proxy completion use the reduced coordinates.

Peg-unplug-side retains $2g+\mathbf{1}_{\text{extracted}}$, where $g$ is the native grasp reward. Initial extraction requires the native condition for successful grasping, including more than 1.5 cm of outward motion, and end-effector distance below 3.5 cm. The omitted terms are the contribution from proximity to the target and the completion override. This differs from the coordinate omissions in Reach and Push and should be interpreted as an omitted completion objective.

### F.3 Miscalibrated targets

Drawer's measured target is shifted along the opening axis so that its displacement is 60% of the native maximum. The reward retains the sum of the native grasp and opening terms, multiplied by 5. Lever's measured target lies on the native lever arc at $0.6\pi/2$, using the native lever radius. Its reward combines grasp and tolerance for proximity to the target through the native Hamacher product. Both constructions preserve the actual simulator goal.

Window's physical opening remains 20 cm. The target used to compute reward is displaced back by 12 cm, producing an 8 cm proxy target. The tolerance for proximity to the target retains the native radius and margin rather than recomputing the margin from the shorter displacement. The proxy combines that tolerance with the native grasp reward.

Door-lock measures vertical distance to the native target plus 5 cm. Its reward is the native approach component plus eight times the modified completion tolerance. The latter uses a 0.5 cm reward radius and 10 cm margin. The native success radius remains 2 cm. These distinct radii are retained explicitly because the inherited general task radius is not the radius used by the lock reward.

MuJoCo Reacher rotates the target coordinates from $(g_x, g_y)$ to $(-g_y, g_x)$. If the fingertip is $p$, its harmful discrepancy is $\|p - g\| - \|p - \text{Rot}_{90^\circ}\, g\|$. The difference is signed. Its negative values are

retained, and the native action cost remains unchanged. The episode horizon is 50 steps, unlike the 1000-step DMC Reacher horizon.

### F.4 Incorrect motion and cost measurements

HalfCheetah uses the torso's horizontal displacement per control step to measure velocity. Its bounded reverse credit is $10\tanh(\max\{-v_x, 0\}/4)$, giving a maximum added reward of 10 per step. Swimmer uses the same family with scale 1 and maximum credit 2.5. Native forward terms and task-specific costs remain present.

Ant underreports the native actuator cost through saturation at scale 0.1. The harmful difference is $c - 0.1\tanh(c/0.1)$. It preserves the sign of the true reward's forward component while making behavior that requires substantial control effort appear less costly.

Sweep measures the object's signed $x$ displacement, with its native goal to the right. Coffee-pull measures the decrease in the object's distance to the visible goal. Both use a 1 cm transition scale and bounded reverse credit with coefficient 8. The added reward is zero for stationary or goal-directed motion and positive for reverse motion. Coordinates are taken from the observed transitions at learner precision.

Pendulum uses the change in pole verticality as progress. Its scale is 0.05 and coefficient 1.5, giving added reward in $[0, 3]$. This construction preserves the native sparse upright reward while introducing an incentive to move away from upright. These functions assign nonnegative credit to reverse motion, generating reward on outward portions of repeated movement. Their accumulated values vary with the policy and are treated as reward errors separately from the supplied potential difference.

Faucet and Handle-pull use native proximity and completion components. Their added terms reward remaining error only when the agent is near or grasping the relevant object. This preserves the prerequisite interaction while reversing the incentive to finish. Their coefficients, 12 and 16, are fixed across runs.

### F.5 Localized false completion

Ball-in-Cup uses the native capture extents after subtracting ball radius. For each horizontal and vertical displacement, tolerance is one inside the capture extent and decreases according to a squared cosine profile over four further extents. The added term rewards lift together with lateral misalignment. It supplies no extra credit inside genuine capture or far below cup height.

DMC Reacher adds a smooth shell of false credit outside the native capture radius $R$. Its support is strictly $R < d < 2R$, and its maximum is 1.25. The credit is zero at both boundaries and inside genuine capture. This construction creates a local alternative to task completion without omitting a spatial coordinate.

### F.6 Verification of reward construction

Environment checks compare native rewards, success, termination, and observations before and after the reward wrapper. Analytic decomposition metadata must sum to the emitted proxy and remain absent from model inputs. Geometry checks verify that the measured target or omitted component differs from the intended objective under reachable states.

The table describes registered reward errors rather than inferred behavior from training curves. Whether a particular policy exploits the incentive is an empirical property measured by native return, success, and task diagnostics. Each result corresponds to the registered construction specified in Table 2.

# G EXPERIMENT DETAILS AND ADDITIONAL RESULTS

## G.1 SUPPLEMENTARY RECOVERY CURVES

For the main comparison on the 22 constructed tasks, we report average intended return over the 600k–1M recovery interval and manipulation success under native success criteria. Final return is evaluated at 1M steps. Results are averaged over six random seeds per method, with each training run constituting one experimental replicate.

Figure 3 presents ten tasks. Figures 7 and 8 add seven manipulation tasks and five further control tasks, completing the 22 constructed scenarios. Each panel compares the same six methods. For display, each run is linearly aligned to the 600k–1M evaluation grid at 10k-step intervals without extrapolation, then smoothed with the binomial kernel with five weights $(1, 4, 6, 4, 1)/16$ and replicated endpoints. Solid lines show the mean across runs.

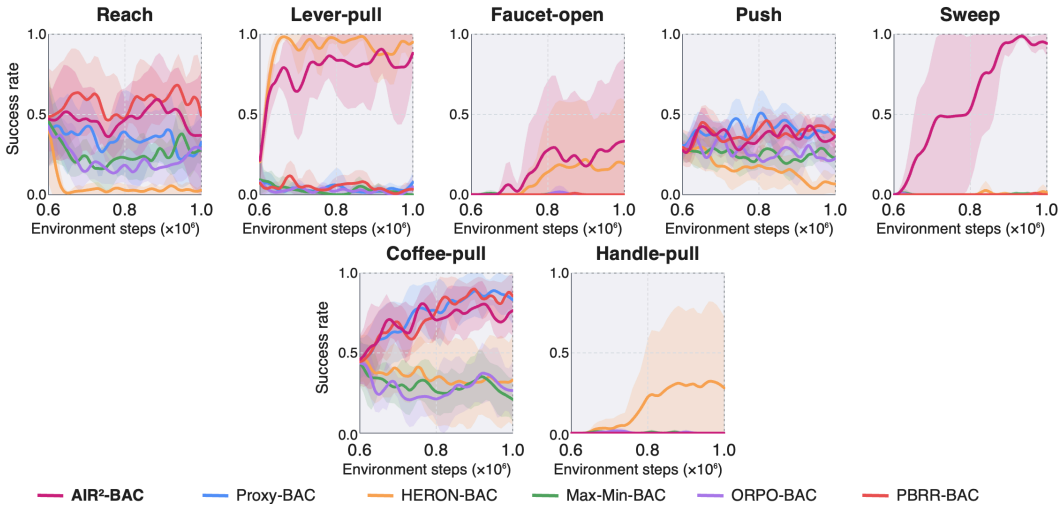



Figure 7: Recovery on seven additional MetaWorld manipulation tasks.

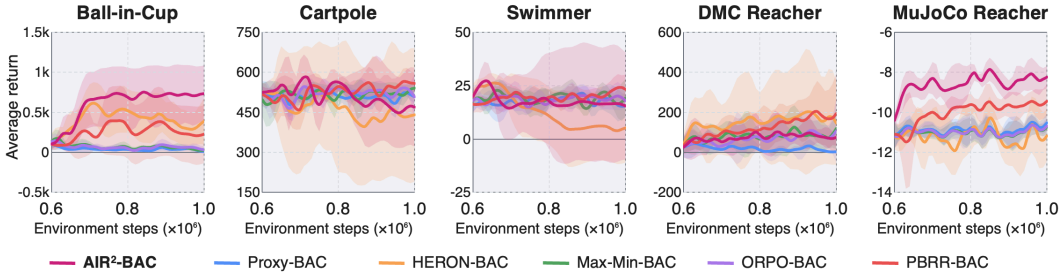



Figure 8: Recovery on the remaining 5 locomotion control tasks.

## G.2 MECHANISM ABLATIONS

We examine potential estimation and feedback design using the BAC continuation setting from Section 5.2.

**Potential estimation.** We remove $\Phi$ while retaining the correction model and fitting feedback. Figure 5 shows the largest losses in average return on HalfCheetah and the manipulation tasks Reach and Sweep. These results support accounting for potential boundaries during reward estimation. The paired return gains are normalized by the full method's average return.

**Behavioral coverage.** Replacing exploration trajectories with an equal number of source policy trajectories weakens recovery on HalfCheetah and Reach. We also compare with HERON using all stored repair feedback. AIR$^2$ retains higher average return on most displayed tasks. These comparisons support collecting informative feedback beyond the source behavior and show that feedback quantity alone does not explain the improvement.

**Temporal feedback.** We fit estimators on $n \in \{16, 32, 64\}$ trajectory pairs from the source policy and vary the horizon set from $\{1000\}$ to $\{250, 1000\}$ or $\{50, 250, 1000\}$. Calibration uses 64 separate pairs at the corresponding horizons. The heatmaps in Figure 6 report the RMSE of predicted outcome differences within trajectory pairs. Evaluation uses common trajectories separate from fitting and calibration, with equal weight across the three behavior groups. At 64 fitting pairs, three horizons reduce prediction error on all four tasks.

**Policy recovery.** We compare one and three horizons at $n \in \{16, 64\}$ fitting pairs. Each variant resumes the same BAC checkpoint at 600k steps and trains for 400k steps with its frozen correction. With 64 pairs, three horizons increase HalfCheetah's average recovery return from 2947.9 to 5863.3. Additional prefixes increase outcome queries for both fitting and calibration. This comparison therefore evaluates temporal information together with feedback availability, using estimators fitted solely on source policy data.

### G.3 PPO EXPERIMENTS ON PUBLISHED BENCHMARKS

Table 3: Average intended return over training on the published PPO benchmarks. Traffic uses $10^3$ return units. Bold marks the largest mean.

| Prevention | | | |
|---|---|---|---|
| Method | Tomato | Pandemic | Traffic |
| Proxy | $5.44 \pm 3.81$ | $-40.02 \pm 2.20$ | $-55.22 \pm 1.20$ |
| HERON | $7.91 \pm 2.95$ | $-10.83 \pm 2.20$ | $-1.78 \pm 1.49$ |
| Max-Min | $7.03 \pm 0.15$ | $-14.68 \pm 3.53$ | $-1.34 \pm 0.02$ |
| ORPO | $7.50 \pm 0.51$ | $-17.03 \pm 8.49$ | $-1.25 \pm 0.03$ |
| PBRR | $7.75 \pm 0.36$ | $-12.24 \pm 8.77$ | $-31.61 \pm 13.28$ |
| AIR$^2$-PPO | $\mathbf{8.27} \pm 2.99$ | $\mathbf{-8.51} \pm 2.45$ | $\mathbf{-0.97} \pm 0.21$ |
| **Recovery** | | | |
| Method | Tomato | Pandemic | Traffic |
| Proxy | $6.78 \pm 4.03$ | $-35.91 \pm 3.32$ | $-56.97 \pm 0.89$ |
| HERON | $7.79 \pm 2.94$ | $\mathbf{-12.53} \pm 5.24$ | $-59.33 \pm 29.12$ |
| Max-Min | $6.68 \pm 3.40$ | $-47.37 \pm 25.94$ | $-55.96 \pm 2.21$ |
| ORPO | $7.63 \pm 2.39$ | $-50.10 \pm 10.56$ | $-61.90 \pm 5.38$ |
| PBRR | $7.07 \pm 3.66$ | $-13.21 \pm 6.88$ | $\mathbf{-44.67} \pm 1.68$ |
| AIR$^2$-PPO | $\mathbf{8.25} \pm 2.02$ | $-12.63 \pm 3.64$ | $-48.32 \pm 24.47$ |

Figure 4 shows the learning curves for the same reported population in both settings. Results are averaged over six random seeds per method. Each run is smoothed with the binomial kernel with five weights $(1, 4, 6, 4, 1)/16$ and replicated endpoints before computing the mean across runs. The curves use intended return of the current policy, and the recovery clock starts at the beginning of the final 40% continuation. Table 3 uses the original unsmoothed curves. Traffic returns in the figure are displayed in thousands, and its prevention inset enlarges the comparison among repair methods.

### G.4 Environment setup

Our evaluation covers 22 continuous control tasks and three published benchmarks. Figure 9 presents the 12 MetaWorld manipulation tasks. The four MuJoCo tasks and six DeepMind Control tasks are shown in Figures 10 and 11, respectively. Figure 12 illustrates the published benchmarks used in the PPO experiments. These views show the underlying environments, whose reward specifications are described in Appendix F and Section 5.

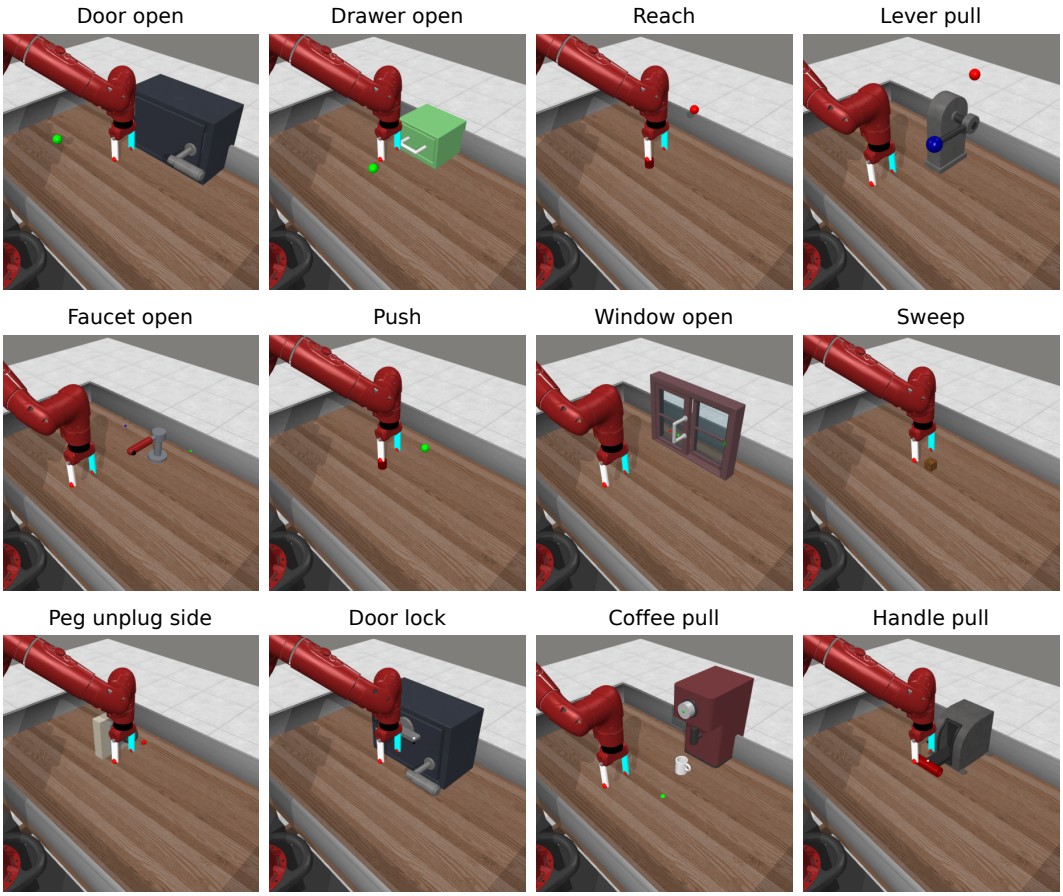



Figure 9: MetaWorld manipulation tasks.

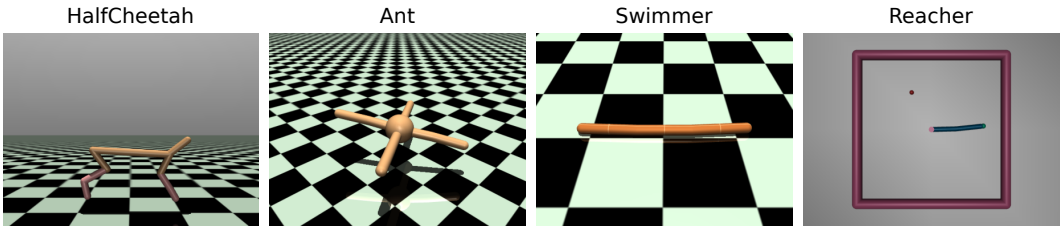



Figure 10: MuJoCo benchmark tasks.

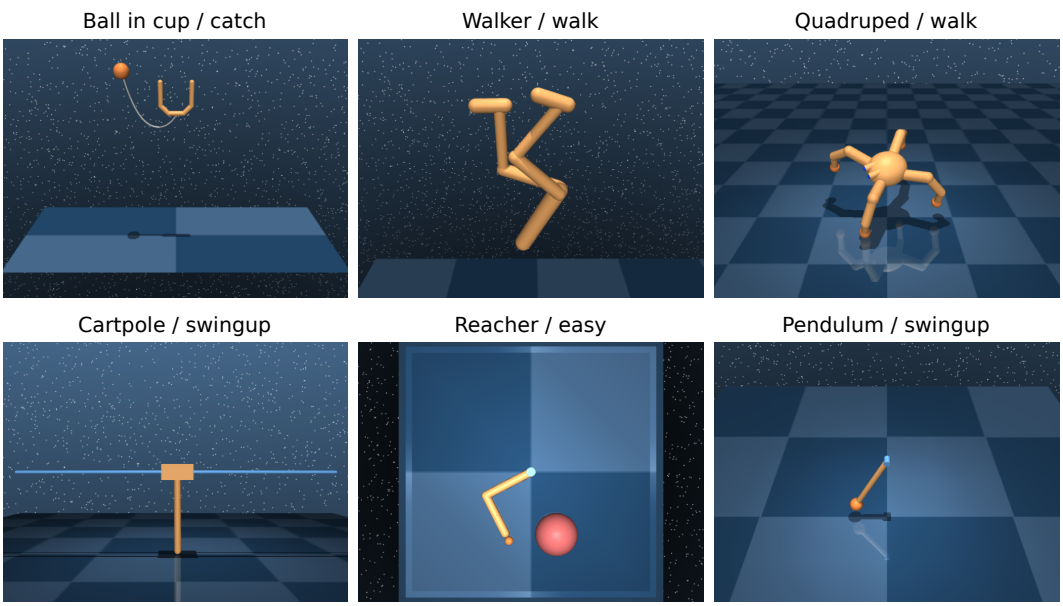



Figure 11: DeepMind Control benchmark tasks.

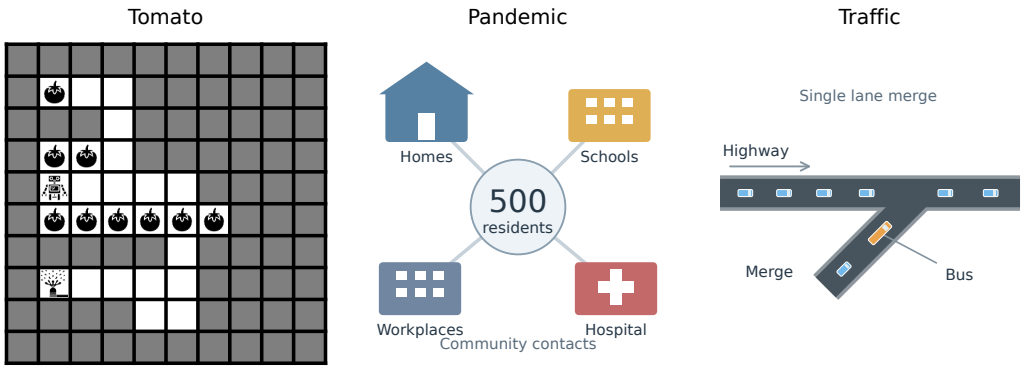



Figure 12: Published benchmark environments. Pandemic and Traffic are shown schematically.

### G.5 Reward repair along policy trajectories

Figures 13 and 14 show reward evolution along individual AIR$^2$ trajectories. Each figure evaluates the reward functions on the same trajectory, with the horizontal axis denoting steps within the episode. The applied correction is $-\hat{b}$, giving the repaired reward $r^p - \hat{b}$. The intended reward $r^\star$ is shown only for diagnosis. **Door-open.** The proxy rewards interaction with the handle but omits the incentive for opening the door. The repaired reward increases as the door opens and remains high after completion, assigning greater reward to the intended behavior. **Sweep.** The proxy gives additional credit for moving the block away from its target. The policy moves the block to the target, where small oscillations still produce proxy reward spikes. Negative corrections attenuate these spikes, reducing the excess reward associated with the incorrect motion measurement.




Figure 13: Reward evolution and policy execution on Door-open.

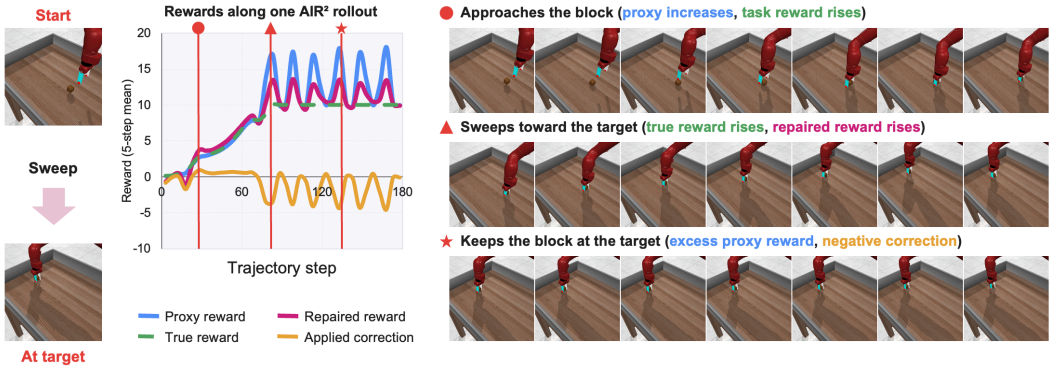



Figure 14: Reward evolution and policy execution on Sweep.

### G.6 Normalized scores

For task $t$ and method $m$, let $\bar{q}_{m,t}$ denote performance averaged over runs and the corresponding training interval, using unsmoothed curves. MetaWorld uses success rates, whereas the other tasks use intended returns. For each task group $\mathcal{T}$, Figure 1 reports

$$S_m = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathrm{clip}_{[0,100]} \left( 100 \frac{\bar{q}_{m,t} - L_t}{U_t - L_t} \right).$$

Thus tasks receive equal weight after normalization and clipping. The reference intervals $[L_t, U_t]$ are $[0, 1]$ for MetaWorld and $[0, 1000]$ for DMC. HalfCheetah and Ant use $[0, 4800]$ and $[0, 6000]$, respectively. Swimmer and MuJoCo Reacher use $[0, 360]$ and $[-120.5, 0]$, respectively. For the published benchmarks, $L_t$ is Proxy's average intended return over the full training horizon. We use $U_t = 18$ for Tomato and $U_t = 0$ for Pandemic and Traffic. All methods share the same reference values, which are also held fixed across prevention and recovery.

# H Additional related work and interpretation

## H.1 Reward misspecification and learning from a supplied reward

Reward hacking arises from objective misspecification when optimizing the available reward favors behavior that conflicts with the intended task (Amodei et al., 2016; Skalse et al., 2022). Corrupted reward channels provide a broader model in which observations of reward can themselves be unreliable (Everitt et al., 2017). Empirical studies document how reward misspecification interacts with optimization strength (Pan et al., 2022; Gao et al., 2023), while studies of reward engineering highlight the difficulty of combining task measurements into a suitable objective (Booth et al., 2023; Knox et al., 2023).

A supplied reward can nevertheless encode useful knowledge. Inverse reward design treats a designed reward as evidence about the designer's intent and reasons about uncertainty in that interpretation (Hadfield-Menell et al., 2017). Our formulation instead uses an observed dense proxy together with sparse trajectory outcomes. The outcomes constrain the accumulated discrepancy between the proxy and intended reward, including potential boundary terms. This information can support a correction while retaining useful dense guidance. Whether it determines the required policy comparisons depends on the admissible correction class and the coverage of the observations, as characterized in Section 3.

## H.2 Reward shaping after the classical invariance result

The potential construction of Ng et al. (1999) remains a foundation for reward transformations that preserve decisions. Subsequent work studies how to learn or adapt useful shaping. Hu et al. (2020) learn to utilize shaping rewards, and Adamczyk et al. (2025) use a learned value estimate as a potential. Forbes et al. (2025) develop shaping that depends on the action and preserves optimality under additional conditions. These developments emphasize both the invariance property and the practical choice of a shaping function.

Our potential has a specific statistical role. Sparse aggregate supervision contains endpoint contributions of any shaping already present in the proxy. Modeling those contributions helps distinguish shaping from a correction that changes policy preferences. The deployed reward retains the proxy and subtracts the selected correction. This use differs from designing an additional shaping reward to accelerate a known objective.

Learning from demonstrations supports reward inference or policy recovery from expert behavior (Ng & Russell, 2000; Abbeel & Ng, 2004; Ho & Ermon, 2016). AIRL separates a reward term and a discounted potential difference in its discriminator (Fu et al., 2018). Its reward disentanglement conclusions depend on the reward structure and dynamics conditions in that work. Our observation equation instead follows from sparse trajectory discrepancies around an observed proxy. The common decomposition motivates the notation, while the supervision and identification questions differ.

## H.3 Reward comparison and robust decisions

EPIC measures reward differences while accounting for reward equivalences (Gleave et al., 2021). STARC studies a general family of standardized reward comparison metrics and connects them to worst-case regret (Skalse et al., 2024). These works motivate assessing rewards by decision consequences rather than only pointwise prediction error. Our seminorm is tied to the declared policy class and retains the cardinal unit fixed by the evaluator.

Skalse et al. (2023) characterize invariances of policy optimization and feedback in reward learning. Rob-ReL connects partially identified reward sets to downstream applications and derives a robust approach to policy comparison (Lazzati & Metelli, 2026). The present formulation concentrates on repairing a supplied dense reward. Its observation operator includes the shaping nuisance and distinguishes the information supplied by temporal queries from the information supplied by new behavior.

Small prediction error under a behavior distribution can coexist with poor optimization performance under a different distribution (Fluri et al., 2025). Related Goodhart effects arise as an imperfect

proxy is optimized more strongly (Karwowski et al., 2024). The ambiguity modulus and policy class make this distributional dependence explicit in our transfer bound.

### H.4 Policy optimization under reward misspecification

ORPO regularizes a policy's departure from a reference occupancy under assumptions on proxy correlation (Laidlaw et al., 2025). Its analysis uses an occupancy divergence to control intended improvement. Max-Min constructs a reward ambiguity set from reference reward moments and a correlation constraint, then optimizes against the least favorable reward in that set (Liu et al., 2026). These methods use a supplied proxy and reference behavior without our sparse outcome supervision. The correlation constraints are modeling information about the intended reward. Our observations supply another kind of information, through aggregate discrepancies. Both approaches depend on coverage of the policies being compared. Complementary approaches modify the optimization objective or the learning process to reduce exploitation of imperfect reward signals (Zhang et al., 2024; Farquhar et al., 2025; Opryshko et al., 2026).

### H.5 Preference learning and residual rewards

Preference learning infers rewards from comparisons of behavior (Christiano et al., 2017). Reward-rational choice provides a framework for relating several feedback modalities to their underlying assumptions (Jeon et al., 2020). HERON converts an ordered hierarchy of feedback signals into preferences and learns a reward model (Bukharin et al., 2025). It also supports renewed feedback collection across multiple stages of policy learning.

Recent reward repair methods learn additive corrections from preference feedback and acquire informative new behavior (Hatgis-Kessell et al., 2026). Residual Reward Models use prior reward knowledge to improve preference-based RL (Cao et al., 2026). These methods establish the value of retaining existing reward information. Our focus is the identification of correction classes modulo shaping and the transfer from sparse outcome fit to policy regret.

Causal confusion can affect the reward learned from preferences when spurious features correlate with the observed labels (Tien et al., 2023). Excluding analytic task diagnostics and proxy values from the correction input limits specific shortcuts in our controlled implementation. It does not replace the need for informative observations under varied behavior.

### H.6 Return decomposition and temporal information

RUDDER redistributes delayed returns through contribution analysis (Arjona-Medina et al., 2019). Randomized return decomposition regresses aggregate outcomes over sampled transitions (Ren et al., 2022). Diaster learns implicit subtrajectory assignments (Lin et al., 2024), while GRD uses a causal generative model of latent Markov rewards (Zhang et al., 2023). These methods provide complementary approaches to temporal credit assignment.

RL with trajectory feedback studies learning without direct observations of transition rewards (Efroni et al., 2021; Zhang et al., 2025). The benefit of feedback on trajectories depends on the feedback model and its information content (Du et al., 2025). Our refinement theorem states an inclusion of admissible explanation sets. It does not imply a uniform improvement in sample complexity for every choice of shorter trajectories. The nuisance representation and statistical dependence of the labels determine how much information is gained.

### H.7 Feedback collection, uncertainty, and policy improvement

Active preference methods choose informative comparisons and account for query redundancy (Sadigh et al., 2017; Bıyık et al., 2024a). Reward learning with Gaussian processes combines nonlinear models with active query selection (Bıyık et al., 2024b). PEBBLE learns from preference feedback while using exploration and replay relabeling (Lee et al., 2021). These connections support treating feedback collection as an integral part of reward learning.

High-confidence policy improvement and baseline bootstrapping study conservative policy changes under statistical uncertainty (Thomas et al., 2015; Laroche et al., 2019). Preference-based variants

extend this perspective to learned rewards (Tse et al., 2025). Constraints relative to a reference policy also support the use of heuristic rewards while seeking improvement under the task objective (Lee et al., 2024). Our primary implementation instead selects among active correction candidates. The transfer theorem makes its guarantee conditional on approximation and selection errors, with identification determining their implications for control.

The confidence analysis for fixed features uses self-normalized estimation (Abbasi-Yadkori et al., 2011). Time-uniform concentration provides additional tools for sequentially acquired data (Howard et al., 2021). Offline RL likewise illustrates the role of coverage and realizability, including single-policy concentrability (Zhan et al., 2022). These statistical tools can support stronger data collection procedures when their assumptions match the implementation. Reliable empirical comparisons complement these analyses through uncertainty estimates and aggregate performance summaries across tasks (Agarwal et al., 2021).